Early Prediction of Student Profiles based on Performance and Gaming Preferences

Gabriel Barata, Sandra Gama, Joaquim Jorge, Senior Member, IEEE, and Daniel Gonçalves

Abstract—State of the art research shows that gamified learning can be used to engage students and help them perform better. However, most studies use a one-size-fits-all approach to gamification, where individual differences and needs are ignored. In a previous study we identified four types of students attending a gamified college course, characterized by different levels of performance, engagement and behavior. In this paper we present a new experiment where we study what data best characterizes each of our student types and explore if this data can be used to predict a student's type early in the course. To this end we used machine-learning algorithms to classify student data from one term and predict the students' type on another term. We identified two sets of relevant features that best describe our types, one containing only performance measurements and another also containing data regarding the students' gaming preferences. Results show that performance alone can be used to predict student type with 79% accuracy by midterm. However, its accuracy improves when paired with gaming data at earlier stages of the course. In this paper we clearly describe our findings and discuss the lessons learned from this experiment.

Index Terms—Gamified learning, Cluster analysis, Student classification, Adaptive learning

1 Introduction

In the last decade, we have witnessed the proliferation of the use of games in learning. This was greatly motivated by the ability of that medium to captivate its users and lead them to endure and strive to complete their goals [1], [2]. Unlike traditional learning materials, games deliver information on demand and within context [3]. This is paramount to prevent players from getting either frustrated or bored [1], [2], which could ultimately lead them to forfeit. Indeed, games have a great potential to engage students and facilitate learning [4], [5], [6]. This is supported by several works, which demonstrate that videogames can successfully be used to motivate students to learn and to improve their learning outcomes [7], [8], [9].

Leveraging on the motivational qualities of games, Gamification uses game-design elements in non-game contexts [10], [11], to engage users into adopting specific behaviors and add value to their experience [12]. For example, gamification has been used to raise fitness and health awareness [13], help in driving instruction [14], improve productivity [15], and promoting loyalty to a brand [16].

Gamification has also been explored as a means to educate, with prominent online services such as Khan Academy [17] and Codecamedy [18] using game elements like points and badges to track user progress and encourage them to learn. Undeniably, research suggests that gamification may significantly increase student activity [19] and performance [20]. However, little has been done to understand how different students learn with gamification and how their particular needs should be addressed.

We have previously gamified a college course where several game elements were added, like experience points and levels, badges, leaderboards, and challenges [21]. Students were more proactive and participate using gamification, and they considered our course to be more interesting and motivating than other non-gamified courses [22].

In an early exploratory study, we analyzed how different students performed in our course and classified them into several types using cluster analysis [23]. We then analyzed how they differed in terms of performance and gaming preferences. We identified four student types: the Achievers, who focused on the achievements and strived to acquire all the available experience points; Regular students, who had above average performance and balanced the achievements with the traditional evaluation components; Halfhearted students, who presented below average performance and seem to have neglected some aspects of the course; and the Underachievers, who had the lowest performance seem to have done just enough to pass the course. While the first two types comprise highly-performing and engaged students, the other two include students that were seemingly disengaged with the course.

In this paper we address an issue that was never explored in any previous research: how can we take advantage of what differentiates students in a gamified learning environment to predict their behavior early in the term? This could potentially be used to customize content and help students with different needs. We used our course as the test bed for a new study, guided by the following research questions:

1. *Were the clusters identified in both experiments the same, i.e., had the same meaning?* This step aimed at verifying the consistency of our clusters between years. If this does not hold, there is nothing to predict.
2. *Is there a subset of relevant features that can be used to predict the student type in this experiment's sample?*

- G. Barata is with the INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal. E-mail: gabriel.barata@ist.utl.pt.
- S. Gama is with the INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal. E-mail: sandra.gama@ist.utl.pt.
- J. Jorge is with the INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal. E-mail: joaquim.jorge@inesc-id.pt.
- D. Gonçalves is with the INESC-ID, Rua Alves Redol, 9, 1000-029, Lisboa, Portugal. E-mail: daniel.goncalves@inesc-id.pt.

This feature-selection process aims at identifying a robust set of relevant features.

3. *Can the relevant feature set be used to predict the students' class in another instance of the course?* This intermediate step would help to further assess the robustness of the clusters across years.
4. *Can student types be predicted by midterm?* This is our main research question.

In this paper we describe how we used cluster analysis, feature selection and classification to answer each of the four questions. We discuss the lessons learned from this study and propose that the use of the information, which can be found by applying our approach to existing learning environments, may be the basis for allowing them to promptly adapt to students with different traits and needs.

2 Background

Serious Games, which are games designed to educate and not necessarily to entertain, have long been used in education with success, with notable gains in terms of motivation, understanding of taught topics, and even performance. Notable examples can be found on distinct subjects, such as programing [24], numerical methods [7], electromagnetism [9] or biology [8], and at several academic levels [7], [25], [26].

Gamification draws on the engagement qualities of good games to encourage students to adopt behaviors that can help them learn better. It can be told apart from Serious Games as it uses game design elements (only) in non-game contexts, instead of full-fledged games [10], [11]. Although there is no formal list of what game elements should be used in gamification, the most consensual seem to be [27], [28]: 1) experience points and levels, serving the main purpose of transmitting feedback and progress; 2) challenges or quests, providing tasks with clear goals, progress assessment and training users for more complex tasks; 3) badges, collectible artifacts that aim at boosting the user's motivation by appealing to her natural desire to collect; and 4) leaderboards, which spur competitiveness and encourage users to continually strive to achieve their desired ranking.

Recent research has focused on how gamification can be used to improve learning. On his book, Sheldon [29] showed how a conventional course could be turned into an exciting game, where students start with an F grade and go all the way up to an A+, by completing challenges and gaining experience points. Domínguez et al. [20] made a comparative study of an e-learning ICT course, where students undertook optional exercises either via a PDF document or a gamified system. In the latter, students were awarded with badges and medals on completion. Students who completed the gamified experience performed better in practical assignments and had higher overall score. However, they appear to have performed poorly on written assignments and participated less on class activities. Cheong et al. [30] used a gamified quiz to evaluate IT undergrad students, where they received points for answering questions and used a leaderboard to compare scores with others. Students self-reported that the quiz helped them perform better and also improved their enjoyment and engagement, but no empirical results were presented.

Haaranen et al. performed another study where they added badges to an evaluation component of a college course [31], which were earned by merit and had no further social meaning. Results show that the addition of badges did not have a significant impact over student performance and behavior, and overall, students were neither engaged nor disengaged by them. In a follow-up study, Hakulinen and Auvinen observed how students with different goal orientations were motivated by badges [32]. The authors divided the student population by goal orientation and collected several measures of student behavior and asked students to provide feedback about the badges via a questionnaire. The authors found "no statistically significant differences in the behavior of the different goal orientation groups regarding badges." However, their attitudes towards the badges varied.

Aguilar et al. studied correlations between college student's perceptions of gamified grading systems and adaptive outcomes associated with gameful course [33], [34] and found those perceptions to be positive and motivating. They observed that "whether students 'like' the grading system is positively related to whether they feel encouraged to work harder". Schutter [35], [36] compared formal measurements of student intrinsic motivation and engagement for a gamified and a non-gamified course on the principles of game design. They concluded that "gameful instruction did not necessarily lead to higher levels of intrinsic motivation or engagement in comparison to traditional teaching methods, and that further improvements to the design and documentation of the course are necessary."

The usage of gamication in education has been somewhat controversial. Indeed, it presents a great potential to shape student behavior and to encourage them to perform better. Most approaches rely on external rewards like badges, but these are prone to decrease the person's intrinsic motivation to perform the task [37], [38] – this is called "overjustification" [39]. Deterding posits that gamified approaches usually miss three ingredients [11]: 1) Meaning – game elements are meaningless unless they are connected to a goal the user has interest in; 2) Mastery, which emerges from providing interesting challenges, clear and varied goals, scaffolded and appropriately paced; and 3) Autonomy, the ability for one to make choices of her own. If used as the main motivator to perform a task, a reward is perceived as control. However, if paired together with goals that are meaningful to the user, more autonomous and internalized behaviors are likely to emerge [38].

Another problem of gamification is its typical reliance on elements that publically display performance. Studies suggests that failure in a public setting can have a negative effect on one's self-esteem and learning performance [40], [41]. However, a recent studies in gamified settings did not support a negative impact by the usage of a leaderboard on the users' intrinsic motivation [42], [33].

Student differentiation and classification has been a hot topic in educational research. Several studies have tried to classify gifted students regarding their achievement and underachievement [43], [44], [45], or to distinguish different learning styles [46], [47]. Machine-learning techniques have also been used to predict student performance and comprehension. Larkey [48] trained Naïve Bayes classifiers and k-nearest-neighbors classifiers to assign scores to manually-graded essays. Pattanasri et al. [49] used Support Vector Machines to predict student comprehension of slides displayed in class, based on self-reported comprehension levels. Minaei-Bidgoli et al. [50] used several classifiers to classify students using logged data in an online learning system and predict their final grade.

Differences in users of gamified services and applications were recently addressed Koivisto and Hamari [51], which studied demographic variations on the perceived benefits of using a gamified fitness service. The authors found that ease of use was negatively influenced by age and gender, with women perceiving more social benefits from gamification, both reciprocal benefits and recognition. On the other hand, social influence was negatively affected by time using the service, and network exposure was predicted by gender and time using the service. The authors also found that perceived playfulness was positively predicted by gender and negatively by time using the service. Time using the service also had a negative effect on enjoyment and usefulness.

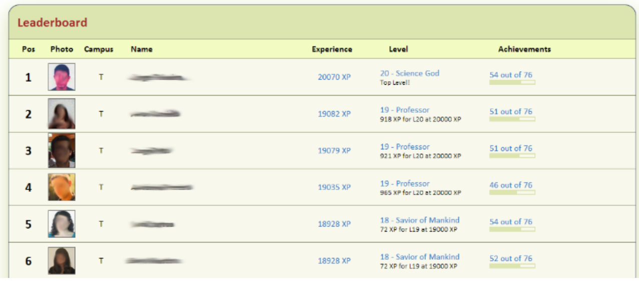
Research shows that we can already predict several aspects of student performance and perception on regular courses. However, a gamified environment presents a whole new learning experience, which is not yet well studied and to which applicability of previous research is questionable. To our knowledge, no prior studies tried to understand how different students of gamified courses can be characterized and if these differences can be used to somehow predict their performance and behavior.

3 Previous Findings

In our previous research we investigated how different types of students performed in a gamified learning environment and how these types were related to the students' gaming preferences [23]. To achieve this we monitored how students of a college course named Multimedia Content Production (MCP) progressed over a term. We then used cluster analysis to identify different progression patterns, which defined four student types. In this section we briefly describe the course and the experiment that led to our student classification model.

3.1 The MCP Course

MCP is a gamified semester-long MSc course, taught yearly at Instituto Superior Técnico, University of Lisbon. The course follows a blended learning model, where students attend live theoretical lessons and lab classes, but also engage in discussion on the course forums, powered by the Moodle platform [52]. Theoretical lectures cover multimedia topics such as capture, editing and production techniques, multimedia standards, copyright and Digital Rights Management. In lab classes, varied concepts and tools are introduced on image, audio and video manipulation, and there are regular assignments as well.



Leaderboard

Pos	Photo	Campus	Name	Experience	Level	Achievements
1		T	[redacted]	20070 XP	20 - Science God Top Level!	54 out of 76
2		T	[redacted]	19082 XP	19 - Professor 918 XP for L20 at 20000 XP	51 out of 76
3		T	[redacted]	19079 XP	19 - Professor 921 XP for L20 at 20000 XP	51 out of 76
4		T	[redacted]	19035 XP	19 - Professor 965 XP for L20 at 20000 XP	46 out of 76
5		T	[redacted]	18928 XP	18 - Savior of Mankind 72 XP for L19 at 19000 XP	54 out of 76
6		T	[redacted]	18928 XP	18 - Savior of Mankind 72 XP for L19 at 19000 XP	52 out of 76

Fig. 1. The MCP leaderboard.

Instead of receiving traditional grades, students earn experience points (XP) in a game-like experience, by undertaking and completing diverse course activities. These include a multimedia presentation (20% of total XP), lab assignments (15%), a final exam in the first instance of the course, which was replaced by regular quizzes in the second year (30%), Skill Tree participation (10%), and a set of collectible achievements (30% plus a 5% extra). These require students to perform specific tasks, such as attending lectures, finding relevant resources related to specific subjects, finding bugs in class slides, or completing challenges, in exchange for XP and badges. Challenges are time limited tasks were students have to produce creative content in response to a specific request from faculty, related to subjects taught in class.

The entry point of our gamified experience is the leaderboard, which displays students sorted by descending order of amount of XP (see Fig. 1). It takes the form of an online webpage that is available from the forums, and that is updated several times a day, to keep student data up to date. Students participate mostly via posts to dedicated threads on forums, which are rated with a score between 0 and 4 by faculty. Student contributions are measured by the sum of their posts' ratings.

Students start with 0XP and earn more by completing course activity. For each 1000XP students increase in experience level. They have to reach level 10 to pass the course, with the top level being 20 (20000XP). Experience levels directly translate to the traditional 20-point grading system used in our university.

There is a special kind of achievement that comprises participating in the MCP Quest, an online treasure hunt where students start from a webpage with a multimedia artifact, which they have to edit and manipulate to find the

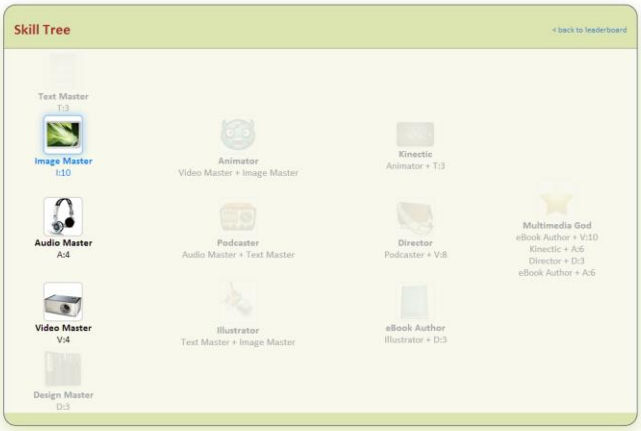


Fig. 2. The MCP Skill Tree.

URL for the next clue of the quest. The amount of XP earned is proportional to the quest level reached and the number of students that actively participate (the more they collaborate, more XP everyone gets).

Apart from the Achievements, there was another game element: the Skill Tree (see Fig. 2). It consists of a precedence tree where each node represented a thematic task, which would earn students XP upon completion. To unlock a new node, the preceding ones had to be completed. This allowed students to earn the maximum grade from this component through different paths, doing more of what they liked or were proficient at.

Course evaluation was identical in both instances of the course, with the exception of the exam. In the second year it was replaced by regular quizzes, occurring usually every other week. By the end of the first year we asked students whether they would prefer to have an exam or quizzes instead, using a Likert scale (1 = Exam, 5 = Quizzes). The majority preferred the quizzes (median: 5, mode: 5).

We have previously shown that our approach is effective at engaging students [21], [22], with them presenting higher levels of participation as compared to previous non-gamified versions of the course, and reporting to be more motivated, interested and learning easier, compared to other regular courses.

3.2 Student Classification

In a previous experiment we identified different types of students based on how they accumulated XP over time [23, 53]. We used Weka [54], a collection of machine learning algorithms for data mining tasks in Java, to perform cluster analysis and group learners by similarities of XP acquisition over time. The algorithm used to achieve this was the Expectation-Maximization (EM) algorithm [55], which does not require the number of clusters to be specified beforehand and works well with small datasets [56]. After identifying the different student clusters, we then observed the average values for several performance and participation measures, as well as the median and modes of their responses to a survey devised by us. This survey inquired the students about their gaming preferences and their classification as a player, according to the BrainHex model [57]. This model characterizes players based on neurobiological responses inherent to playing games, and it builds on the popular and validated Demographic Game Design 1 (DGD1) model [58].

We identified four clusters, which we encoded as four student types with different levels of performance, participation and engagement with the course. First we had the *Achievers*, which presented the highest XP accumulation curve, with the steepest slopes. They exceled on all components of the course but they were a minority, representing only 13% of all student population. The second cluster had above average overall performance and their weakest points regarded some components such as the Skill Tree, where they have average performance. Because these students represented the bulk of the population, around 43%, we named them the *Regular*. Both Achievers and Regular students were highly participative.

The *Halfhearted* students composed 24% of the student population and they were represented by a below average overall performance, neglecting a few components such as the Skill Tree, the Exam and the Quizzes. These students were not particularly participative, but they managed to explore and complete a moderate amount of achievements. Finally, the Underachievers, who comprised around 20% of the population, presented the lowest XP accumulation curve, with fewer pronounced slopes. They had the poorest performance and participation levels on all components and they seemingly only did enough to pass the course.

Students' responses to the gaming survey did not reveal any significant differences among clusters regarding their playing habits and preferences. However, interesting patterns emerged regarding their classification according to the BrainHex player mode. Achievers were predominantly classified as the Socializer (29%) and the Mastermind (29%), Regular students as BrainHex's Achiever (26%) and Mastermind (26%), Halfhearted as Conquerors (67%) and Underachievers as Conquerors (36%) and Seekers (27%). These patterns suggested that our gamified experience might be more appealing to players that enjoy social experiences, using strategy to make efficient decisions, and to collect artifacts and achieve long term goals.

4 Study and Experiments

We performed a new multi-step study to determine if the students' type could be predicted by midterm, based on the particularities that characterize each student. A new sample was used for the new experiment, which consisted of the enrolled students in the course instance following that described in the previous section. We had 76 students, of which 9 were female, and a large majority of which had finished their undergraduate computer science degree on the previous year. The study was designed in four steps, each portrayed by a research question:

Q1. *Were the clusters identified in both experiments the same, i.e., had the same meaning?* Used to validate our clustering model. If this does not hold, there would be nothing to predict.

Q2. *Is there a subset of relevant features that can be used to predict the student type in this experiment's sample?* This feature-selection process aims at identifying a robust set of relevant features that best discriminate our data. These will be used to train classifiers.

Q3. *Can the relevant feature set be used to predict the students' class in another instance of the course?* This intermediate step would help assessing the robustness of the set's predictive power across years.

Q4. *Can student types be predicted by midterm?* This is our main research question.

We used data from both years, where students were already classified, and compared performance and participation data to assess type consistency and answer Q1. Then, we used a process called feature selection to identify relevant features that could discriminate our students, and thus answer Q2. We proceeded to plot student performance and participation measurements in several points in time for both years, and used data from one year to train classifiers and data from another to test them, to assess whether the students' type could be predicted by the end of the term and by midterm, thus answering Q3 and Q4.

In this section we will describe in detail how each of the four steps were performed and the inherent results.

4.1 Cluster Consistency

This step aimed at answering the first questions. To this end, we performed a second experiment, where we repeated the procedure described in section 3.2, but with the new batch of students. The course lasted for 138 days, but the first nine were excluded from the analysis, during which there were no activity and some students were not fully enrolled in the course. The same criteria was used in the previous experiment to prevent clustering algorithms from overweighting the first days, where there were no significant activity [23]. Like in the previous experiment, our variables did not fit a normal distribution. We checked for differences between clusters using a Kruskal-Wallis test, with post hoc Mann-Whitney's U tests and Bonferroni correction, with a level of significant of 0.8%. Significant differences are reported in Table 1.

Like the year before, cluster analysis revealed four distinct clusters, with similar XP accumulation curves (see Fig. 3) and levels of participation and performance (see Table 1) to those of the first experiment. Thus, clusters retained the same names between experiments. As seen in Fig. 3, the Achievers present the highest XP accumulation curve, with the steepest slopes, again excelling on all aspects of the course and being the most participative. Regular were again represented by an above average overall performance and participation levels, lagging behind in the Skill Tree and MCP Quest in comparison to the Achievers. The Halfhearted students had a below average XP accrual, and performed worse that the Achievers and Regular on the Skill Tree, MCP Quest, the quizzes and multimedia presentation. Again, Underachievers had the lowest performance and participation, and seem to have done just enough to pass the course.

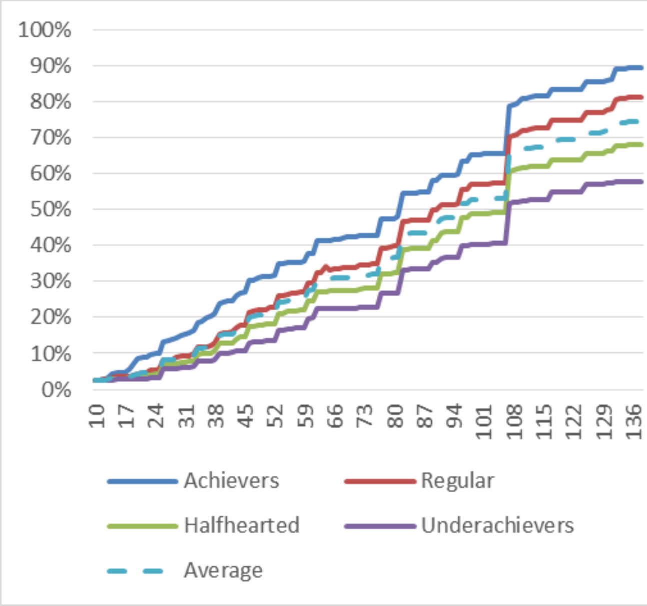


Fig. 3. XP accumulation curves in the new experiment.

In this experiment, we had 11 Achievers (14.5% vs. 12.9% of 1st experiment), 29 Regular students (38.2% vs. 42.6% of 1st experiment), 23 Halfhearted (30.3% vs. 24.1% of 1st experiment), and 13 Underachievers (17.1% vs. 20.4% of 1st experiment). These numbers suggest that clusters on both experiments present comparable proportions, with a minor increase in size of the Achievers and Halfhearted, and slight decrease of the Regular and Underachievers.

We collected gaming data from students via a questionnaire at the beginning of the course and obtained 75 replies. Again, we did not observe any significant differences between student types, but BrainHex classification patterns diverged in comparison to the previous year. Our clusters appeared to be more homogeneous regarding this subject, with the mode being the Conqueror for the Achievers, Regular and Halfhearted, covering 30%, 32% and 50% of the respective populations. For the Underachievers, the modes were the Conqueror and the Seeker, each representing 23% of the population.

Given that our clusters were determined by performance over time, we considered that the similarity between proportions, performance and participation meas-

TABLE 1
CLUSTER PERFORMANCE DATA FROM THE NEW EXPERIMENT. GREEN AND RED DENOTE THE HIGHEST AND LOWEST LEVELS.

Property	(A) Achievers	(B) Regular	(C) Halfhearted	(D) Underachievers	Student Average	Significant Differences ($p < 0.008$)
Quizzes Grade (%)	74.05	72.4	66.96	65.58	69.83	none
Labs Grade (%)	94.62	88.59	85.29	74.49	86.05	(A, D)
Presentation Grade (%)	83.64	81.12	71.81	72.58	77.21	none
Final Grade (%)	89.53	81.4	68.01	57.84	74.5	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
Attendance (%)	92.64	94.42	79.5	78.02	86.84	(A, D), (B, C), (B, D)
Posts (#)	70	47.17	24.57	12.62	37.72	(A, C), (A, D), (B, C), (B, D), (C, D)
First Posts (#)	4.27	3.41	1.04	0.54	2.33	(A, D), (B, D)
Reply Posts (#)	65.73	43.76	23.52	12.08	35.39	(A, C), (A, D), (B, C), (B, D), (C, D)
Rated Posts (#)	43.64	28.97	15.83	7.62	23.46	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
Mean Rate	3.06	2.95	3.01	2.88	2.97	none
Challenge Posts (#)	17.45	17.38	12	7.54	14.08	(A, C), (A, D), (B, C), (B, D)
XP from Challenges (%)	100	95.07	73.91	42.12	80.33	(A, C), (A, D), (B, C), (B, D), (C, D)
Theoretical Challenge Posts (#)	10.09	9.93	6.61	4.23	7.97	(A, C), (A, D), (B, C), (B, D)
XP from Theoretical Challenges (%)	100	96.55	69.57	41.03	79.39	(A, C), (A, D), (B, C), (B, D)
Lab Challenge Posts (#)	7.36	7.45	5.39	3.31	6.11	(A, D), (B, D)
XP from Lab Challenges (%)	100	93.1	79.71	43.59	81.58	(A, D), (B, D), (C, D)
Skill Tree Posts (#)	19.82	12.03	5.91	1.69	9.54	(A, C), (A, D), (B, C), (B, D), (C, D)
XP from Skill Tree (%)	75.68	47.07	23.7	5.96	37.11	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
Explored Skill Tree Nodes (#)	9	5.86	3.09	0.77	4.61	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
MCP Quest Posts (#)	8.45	3.66	1.91	0.31	3.25	(A, C), (A, D), (B, D)
XP from MCP Quest (%)	90.91	82.76	47.83	7.69	60.53	(A, D), (B, D)
Badges (#)	46.27	38.55	27.65	21.38	33.43	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
XP from Achievements (%)	97.95	89.9	64.84	43.37	75.52	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
Completed Achievements (#)	14.45	10.62	5.87	3.46	8.51	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)
Explored Achievements (#)	20.55	18.21	14.96	12.54	16.59	(A, B), (A, C), (A, D), (B, C), (B, D), (C, D)

urements amid experiments suggests Q1 can be affirmative, although it does not prove it. The ideal way to verify Q1 would be to assess concordance between the classifications of the models, resultant from cluster analysis on both course instances. However, this is not feasible owing to the different number of days and differences in the evaluation criteria between experiments. As alternative, we assumed Q1 to be true, and proceeded to verify Q2 and then Q3. If Q3 holds true, then Q1 is more likely to be true as well.

4.2 Relevant Feature Selection

Our second question consisted on whether a subset of relevant features can be found that best describes our students. This problem is solved through a process named Feature (or Attribute) Selection, which uses a search algorithm to search through possible features subsets and another to evaluate the selected attributes by running the model on the subset. The one with the best descriptive power is selected.

We performed features selection by using the "Select Attributes" feature of Weka. For attribute evaluation we used the Correlation based Feature Selection (CFS) algorithm [59], with the search algorithm being the Best First.

Given that we had more information available in the second year (i.e., lager number of cases), we used student data from that experiment to perform feature selection. We considered two sources of data: a) student performance data, which consisted of automatically collected performance measures for all the 76 students (described below); and b) student gaming preferences and BrainHex classification, which we collected via a questionnaire at the beginning of the course from 75 students. We performed feature selection on the two datasets in order to assess whether gaming data, which can be collected before the course starts, would be considered relevant or not.

Our Performance dataset consisted of several distinct performance measurements, which included those described in Table 1 plus the number of badges acquired per student for all of the achievements. Feature selection on this dataset yielded, with a merit of 0.686, the following seven features:

- Current Grade (%) – the total amount of XP, in percentage, accumulated so far.
- Rated Posts (#) – the number of rated posts.
- Skill Tree Posts (#) – the number of Skill Tree posts.
- Badges (#) – the number of collected badges.
- XP from Achievements (%) – the amount of XP earned from Achievements, in percentage.
- Completed Achievements (#) – the number of completed Achievements.
- [A] Artist – the number of badges acquired in the Artist achievement. This was a three-level achievement that required students to make four, six and 12 posts with the top rating.

TABLE 2

CROSS-VALIDATION ACCURACY AND RELIABILITY RESULTS.

Classifiers	Participation Dataset			Gaming Dataset		
	AUC	Kappa	Acc.	AUC	Kappa	Acc.
BayesNet	0.90	0.63	73.68%	**0.91**	**0.70**	**78.67%**
SimpleLogistic	**0.90**	**0.67**	**76.32%**	0.89	0.58	70.67%
SMO	0.88	0.61	72.37%	0.87	0.56	69.33%
IB1	0.71	0.43	59.21%	0.74	0.49	64.00%
J48	0.84	0.65	75.00%	0.85	0.59	70.67%

The second dataset, which we named Gaming dataset, contained data from the gaming characterization questionnaires, which included the BrainHex classes, plus the same data from the Performance dataset for the 67 students that answer the questionnaire. Feature selection returned the following attributes, with a merit of 0.687:

- Current Grade (%)
- Challenge Posts (#) – number of challenge posts.
- Skill Tree Posts (#)
- Badges (#)
- XP from Achievements (%)
- Completed Achievements (#)
- [A] Artist
- BrainHex Main – the student's BrainHex Main class.

Attribute selection from both datasets produced comprehensive and concise subsets of features, with high levels of merit, which is a good predictor of accuracy [59]. Six features are common to both feature sets, which include the student's current grade, number of Skill Tree Posts, badges acquired in the Artist achievement, completed achievements, total number of badges, and amount of XP earned from achievements. The number of rated posts were considered relevant only for the performance dataset whereas the number of posts made in challenges and the student's BrainHex main class were only discriminant in the gaming dataset. These findings seems to support a positive response to our second question: we found not only one but two subsets of relevant features, one with performance data only and another with gaming data too.

4.3 Classifier Selection

We examined a set of candidate classification algorithms to later assess the predictive capabilities of our feature sets. The assessment consisted of feeding both the aforementioned datasets to the classification algorithm and performing ten folds cross-validation, using Weka. The resulting performance is the average of the ten classifiers. We evaluated five different classifiers that seem to best fit our data. These algorithms were available in Weka. They were:

- BayesNet – a Bayesian Network.
- SimpleLogistic – a classifier for Logistic Regression models.
- SMO – a sequential minimal optimization algorithm for training a support vector classifier.
- IB1 – a nearest-neighbor classifier.
- J48 – a classifier based on C4.5 decision trees.

We then compared all the algorithms regarding their reliability using the Area under the ROC curve and the Kappa statistic. Kappa (K) measures the agreement beyond that expected by chance [60]. The magnitude of Kappa might be interpreted into difference levels of strength of agreement: poor (≤0), slight (≤0.2), fair (≤0.4), moderate (≤0.6), substantial (≤0.8), almost perfect (≤1) [61]. The Receiver Operating Characteristic (ROC) curve is another way to summarize classifier performance, by plotting true positive rate against false positive rate [62]. The

Area Under the [ROC] Curve (AUC) is an accepted performance metric [63], where a value of 1 (100%) means all positive examples were correctly classified and no negative were classified as positive, and 0.5 (50%) means that there were as many positive examples correctly classified as negative examples misclassified, which may be compared to random guessing and has similar meaning to K = 0.

We selected the classifiers with AUC of approximately 0.9 and above, which subjectively indicates excellent accuracy [64], and with at least moderated agreement (K>0.4). The results from cross-validation can be found in Table 2. Our criteria led us to select BayesNet and SimpleLogistic as the classifiers, which seem to be the most reliable for both datasets. The high accuracy, AUC and Kappa levels for both datasets support a positive response to our second question, further confirming that the selected features in the previous step have a high predictive power.

To summarize, we used the feature sets identified in the previous section to evaluate several classifiers, using cross-validation. The classifiers based on Bayesian Networks and Logistic Regression presented the best results.

4.4 Inter-Year Prediction

Our last two questions (Q3 and Q4) aimed at verifying if we could predict the student's type from another year, not only by the end of the course but also early on. We trained models with data from the second year, using the classifiers identified in the previous step, and tested them with data from the first experiment. We used both the performance and the gaming feature sets, in four points in time: 1) five weeks, roughly one month of classes (~25-28% of total time span), 2) seven weeks, around one month and half of classes (~35-38%), 3) nine weeks, around midterm (~45-50%, and 4) end of last week of the course (100%). The fourth time point was considered mainly to answer Q3 whereas the other three served to assess how early and how well we could predict a student's type, and thus answer Q4. For each point in time, features were computed based on how students performed until then. For example, Current Grade represents the total amount of XP accumulated by then, and Challenge Posts represents the amount of posts made on challenges up until that date.

Our data presented a few particularities that required special attention. Firstly, both years presented a small amount of cases, hence the choice of using the second year – the one with the larger number – as the training set. Secondly, both datasets were imbalanced, with some clusters being two to three times larger than others. To deal with this limitation we considered three options: 1) randomly under-sampling the largest clusters, 2) randomly over-sampling the smaller clusters, and 3) Ensemble Learning. We excluded the first two because the former can potentially exclude important data and the latter can lead to overfitting [65]. We opted for Ensemble Learning, which consists of combining several classifiers to improve prediction accuracy. We tested two forms of Ensemble Learning: Voting, which combines the probability estimates of several classifiers; and Stacking, which uses a meta-classifier to learn from the predictions of the other classifiers. Therefore, besides predicting with BayesNet and SimpleLogistic,

TABLE 3
INTER-YEAR PREDICTION PERFORMANCE FOR THE PERFORMANCE DATASET.

Classifiers	After 5 Weeks			After 7 Weeks			After 9 Weeks			End of the course		
	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.
BayesNet	0.66	0.01	21.05%	0.84	**0.50**	**62.96%**	**0.95**	**0.72**	**79.63%**	0.96	0.74	81.48%
SimpleLogistic	0.70	**0.19**	**31.58%**	0.61	0.17	29.63%	0.87	0.37	51.85%	0.93	0.66	75.93%
Voting: Average of Probabilities	0.66	0.18	**31.58%**	0.85	0.19	31.48%	0.93	0.62	72.22%	0.95	**0.77**	**83.33%**
Voting: Product of Probabilities	**0.79**	0.18	**31.58%**	0.70	0.19	31.48%	0.94	0.69	77.78%	**0.97**	**0.77**	**83.33%**
Stacking: LogisticRegression	0.64	0.00	17.11%	**0.88**	0.00	20.37%	0.94	0.61	72.22%	0.95	0.77	**83.33%**

TABLE 4
INTER-YEAR PREDICTION PERFORMANCE FOR THE GAMING DATASET.

Classifiers	After 5 Weeks			After 7 Weeks			After 9 Weeks			End of the course		
	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.
BayesNet	0.72	0.21	39.62%	0.89	0.45	58.49%	**0.94**	**0.62**	**71.70%**	0.95	0.64	73.58%
SimpleLogistic	**0.78**	0.26	43.40%	0.92	**0.53**	**66.04%**	0.89	0.41	54.72%	0.89	0.50	64.15%
Voting: Average of Probabilities	0.76	0.24	41.51%	0.91	0.53	**66.04%**	0.93	0.47	60.38%	0.92	**0.66**	**75.47%**
Voting: Product of Probabilities	0.74	0.24	41.51%	**0.93**	**0.53**	**66.04%**	0.93	0.54	66.04%	**0.96**	**0.66**	**75.47%**
Stacking: LogisticRegression	0.68	**0.31**	**47.17%**	0.89	**0.53**	**66.04%**	0.94	0.47	60.38%	0.93	0.64	73.58%

TABLE 5
INTER-YEAR PREDICTION PERFORMANCE USING ACCUMULATED XP ONLY.

Classifiers	After 5 Weeks			After 7 Weeks			After 9 Weeks			End of the course		
	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.	AUC	Kappa	Acc.
BayesNet	0.48	**-0.02**	**18.42%**	0.76	**0.50**	**62.96%**	0.62	**0.08**	**27.78%**	**0.94**	**0.75**	**81.48%**
SimpleLogistic	0.65	-0.05	17.11%	0.86	0.47	61.11%	**0.76**	**0.08**	**27.78%**	0.93	0.67	75.93%
Voting: Average of Probabilities	0.64	**-0.02**	**18.42%**	0.86	**0.50**	**62.96%**	**0.77**	**0.08**	**27.78%**	0.93	**0.75**	**81.48%**
Voting: Product of Probabilities	**0.66**	**-0.02**	**18.42%**	**0.87**	0.49	**62.96%**	**0.77**	**0.08**	**27.78%**	**0.94**	**0.75**	**81.48%**
Stacking: LogisticRegression	0.48	-0.05	17.11%	0.83	0.47	61.11%	0.68	0.07	**27.78%**	0.93	0.67	75.93%

we also tested with Voting and Stacking with these two classifiers. For Voting, we tested combining both the average and the product of the probabilities, and for Stacking we used a meta-classifier based on Linear Regression.

Prediction performance for all classifiers, for both feature sets, is depicted in Table 3 and Table 4. We have also included the results for baseline classifiers, where only a single feature was used, the amount of XP accumulate so far (Current Grade) (see Table 5).

Results show that the three feature sets can be used to predict the student type at different points in time. After five weeks of classes, up to 47.17% of the students' type could be predicted using Stacking in the gaming dataset, although with low levels of AUC and Kappa, which suggests that these predictions were not much better than chance. SimpleLogistic had slightly lower accuracy (43.40%), but presented a higher AUC level. All models performed better than the baseline for gaming dataset. For this point in time, the performance dataset presented low accuracy, AUC and Kappa for every classifier, performing only slightly better than the baseline models.

After seven weeks, prediction accuracies increased. In the performance dataset, the BayesNet classifier presented the best performance, with a considerable value of AUC (0.84), moderate agreement (K = 0.5) and 62.96% accuracy. All other classifiers performed poorly. On the other hand, for the gaming dataset, all classifiers had a high accuracy rate and AUC levels, with the best results being presented by voting with product of probabilities, with a classification rate of 66.04% (AUC = 0.93 and K = 0.53). Interestingly, the baseline classifiers presented a good performance for this point in time. The BaysNet classifier performed as good as using accumulated XP alone as it did using performance features. However, baseline voting and stacking classifiers fell behind those using gaming features by around 3 to four percentage points.

After nine weeks, most classifiers on the performance dataset started to perform better as compared to the gaming dataset. While with the former, up to 79.63% of the students were correctly classified using BayesNet (AUC = 0.95, K = 0.72), in the latter only 71.70% were matched, also using the same classifier (AUC = 0.94, K = 0.62). Surprisingly, baseline classifiers performed poorly in this time stamp, only correctly classifying 27.78% of the students (AUC = 0.77, K = 0.08). By the end of the course, up to 83.33% of the students were correctly classified using the performance dataset, with both Stacking and Voting (AUC ≥ 0.95, K = 0.77). The baseline classifiers had the second best performance, correctly classifying 81.48% of them using BayesNet and Voting (AUC = 0.94, K = 0.75). Up to 75.47% were correctly classified using the gaming dataset, with both versions of Voting (AUC ≥ 0.92, K = 0.66).

The high prediction accuracy by the end of the course seems to support our assumption that clusters were consistent between both years (Q1), and it suggests that Q3 is also positive. We observed that with the features from the performance dataset we could correctly predict 62.96% of students' type with moderate agreement as early as of seven weeks. This prediction rate increases to 66.04% for the same milestone with the gaming dataset. By midterm, we could correctly predict the student's type with 79.63% and 71.70% accuracy, using the performance and gaming datasets respectively, which suggest that the answer to Q4 is affirmative.

To summarize, we have plotted, for both years, the two feature sets identified in section 4.2 into four points in time, roughly ¼, ⅓ and ½ of the semester, and by the end of the term. We then trained classifiers with data from the second year, which had a larger population, and tested with data from the first year. Results show that the gaming dataset provided a better performance during the first two milestones, but from midterm on, the performance dataset was more discriminant of the students' type.

5 Discussion

With this study we wanted to ascertain whether or not student differences regarding performance and gaming preferences could be used to identify their type by midterm. To attain this, we conducted a new experiment to answer four research questions, which we address in this section.

Q1) Were the clusters identified in both experiments the same, i.e., had the same meaning? Yes. The main goal of this question was to assess cross-years cluster integrity, i.e., if all clusters meant the same in both years. We performed student clustering based on XP accrual on our second year and compared several performance metrics to those of the first year. Clusters appeared to be consistent with only slight divergences regarding their classification according to the BrainHex model. Given that our clusters are based on performance accrual, we assumed Q1 to be true and proceeded to answer the other questions. We sought for further validate Q1 by answering Q3 in a later step.

Q2) Is there a subset of relevant features that can be used to predict the student type in this experiment's sample? Yes. We performed feature selection on two datasets from our second year, one containing only performance measurements (the performance dataset), and another containing performance data and also the students' classification according to the BrainHex model (the gaming dataset). Ten-fold cross-validated classification showed that both could be used to correctly classify more than 75% of the students' type, which allows us to answer Q2 with a "yes".

Q3) Can the relevant feature set be used to predict the students' class in another instance of the course? Inter-year prediction using both feature sets on four different milestones revealed that accuracy grows with time. Because we were dealing with a classification problem that comprises four categories, we considered 60% to be reasonable minimum acceptable accuracy rate. By the end of the course we could correctly predict 83.33% of the students' type using performance features, 75.47% using the gaming features, and 81.48% with the baseline classifier. This answers Q3 affirmatively and further supports Q1's answer to be positive as well. We could correctly predict more that 60% of the students' type, using both sets of features, starting from the seventh week. Gaming features presented a 3% advantage over performance features alone, whereas these did not yield better results than the baseline feature. By the end of the first five weeks, only gaming features could be used to predict 47.17%, with only a fair level of agreement. By midterm, we could predict 71.70% and 79.63% using the gaming and performance feature sets respectively with substantial agreement, which answers with a "yes" our fourth question: *Q4) Can student types be predicted by midterm?*

5.1 Implications for Research

We learned a few lessons from this experiment which we hope can help people do better-informed decisions on future research. Our results suggest that, of the two feature sets we explored, one had better predictive power during the first weeks than the other. It is the case of the gaming features, which included several performance metrics and also the students' classification using the BrainHex player model. Classifiers using these features performed better on the five- and seven-week milestones and worse on the nine-week and end-of-the-course marks, as compared to the feature set containing performance metrics only. This suggests that the BrainHex classification has some power to predict student behavior and performance. Because it can be measured when the course starts, it appears to be an important asset during the first weeks, where student performance data is yet scarce. By midterm, student performance is more well-defined and consistent, and is better portrayed by automatically collected performance metrics, whose predictive ability outperforms that of the gaming features. This is further corroborated by the fact that, by the end of the course, classifiers using these features actually performed worse than using accumulated XP only. We believe this matter should be subject to further research.

Dealing with imbalanced datasets in classification is problematic, especially with a small number of cases. In an attempt to overcome this barrier, we used two methods of Ensemble Learning: Voting and Stacking. Neither of them presented a substantial boost in comparison to the solo classifiers, with the exception of the five-week milestone in the gaming dataset, where Stacking BayesNet and SimpleLogistic together presented the best performance.

We are not certain on whether the clusters and feature sets identified in this experiment can or should be generalized to other studies and gamified learning environments. For now they might serve as a starting point for future research, although this subject should be further investigated. However, we believe that one of our major contributions is the approach presented in this paper, which provides the means that may enable a gamified learning environment adapt to different student needs and traits.

5.2 A New Approach to Gamified Learning

The approach here proposed consists of characterizing students in a gamified learning environment, using diverse sources of data, and adequately training statistical models to predict their behavior and/or performance. Although it was only tested within our course setup and thus was not validated, we believe it can easily be replicated in and adapted to other gamified learning contexts.

This approach has three main assumptions:

1. Student data must be from at least two instances of the course, for a considerable amount of students in each instance (above 50). The nature of the data in our experiment comprised both performance metrics and classification according to a player model, but other sources should be considered as well, such as formal measurements of student engagement or classification according to different learning styles. Let's call this the *student characterization data*.
2. It must be possible to sample a significant part the student data regularly in the course.
3. There must be a single measurement of progress that can be plotted over time, which in our case was XP accrual. Let's call this the *student progression data*.

Our approach leverages on student data to identify different student categories that code different performance and behavior patterns. Classifier algorithms are then used to create models capable of predicting a student's type in another year. Our approach consists of two phases: Student Characterization and Prediction phase.

5.2.1 Student Characterization Phase

This phase aims at identifying what data distinguishes one student from another and then characterizing them accordingly. It comprises of two steps:

1. *Student clustering:* clustering analysis should be performed on the *student progression data*; the Expectation-Maximization algorithm is preferable if samples are small. Each case should represent a student and the performance measurements for each day should be regarded as attributes. The resulting clusters encode different progression patterns.
2. *Inter-cluster analysis:* Descriptive statistics and tests to analyze differences between the clusters' means should be performed for the *student characterization data*. This will reveal which metrics best characterizes each cluster, and to what degree.

This phase should be performed for at least two instances of the course (i.e. two terms).

5.2.2 Student Prediction Phase

The main goal of this phase is to develop a statistical model capable of using data from previous instances of the course to predict the students' type, as identified in the previous phase. This phase has the following steps:

1. *Cluster consistency verification:* clusters observed in the previous phase must be consistent across instances of the course, i.e., they must represent roughly the same student traits. A comparison of the *student characterization data* between clusters must be performed here. If the number of attributes considered in the first step of the first phase is the same in all instances, this step can be extended. The model resultant from cluster analysis from one instance of the course can be tested with *student progression data* from another year, and classification results can be compared
2. *Feature selection:* relevant features from *student characterization data* must be selected to train classifiers. This step is particularly important to prevent the model from overfitting data. This can be done by performing feature selection using Correlation based Feature Selection. Let's call the result *relevant*

student characterization data.

3. *Classifier training:* classification algorithms must be used on *relevant student characterization data* from one or more instances to train a statistical model, capable of predicting the students' cluster. The selection of the algorithms can be done by validating their classification capabilities with cross-validation. Classification accuracy, ROC Area and Kappa statistic should be considered to evaluated classifier performance and reliability. Models should be trained using the best performing algorithms.
4. *Prediction evaluation:* here we validate the models trained in the previous step, by testing them with data from one or more instances of the course, not yet used for training. A compromise must be found between prediction accuracy and reliability to select the best model. If additional samples of *relevant student characterization data* are available from other points in time, these should also be processed.

The second phase yields a statistical model which can be fed with sampled *student characterization data* from a new instance of the course, in a particular moment, to predict what cluster each student belongs to.

We believe that this study can help develop the concept of adaptive gamified learning. We envision a gamified learning environment that makes use of our approach to help predict student behavior and performance early in the course, using both performance and gaming data, and adapt to them in near real-time or on-demand. Such a system could constantly monitor student activity and feed it to a statistical model, which would classify students according to their expected behavior. This information could be used to build and update meaningful progress visualizations tools, which would provide valuable feedback to both students and instructors alike. It could also be used to adapt content and take specific measures to help and guide students at risk and with different needs, which could be either triggered by faculty or pre-programmed. For example, in the case of our course, an Underachiever or a Half-hearted student could be gently reminded of opportunities to gain additional XP, and faculty should be automatically warned once these fell below a certain performance thresholds, so that adequate measures could be promptly adopted. In the same way, an (over)Achiever that is crowding a specific thread and preventing others from participating could be automatically restrained in a non-punitive way, thus allowing others to post under less pressure.

5.3 Study Limitations

Our study has five main limitations. The first one concerns the small sample size used on both years, which might have affected every step of our analysis. The second limitation relates to the difficulty in verifying cluster consistency between years, which is caused by two restrictions: a) there are uncontrolled variables between experiments, such as different number of students, differences in course materials, and the replacement of the exams for the regular quizzes; b) given that the number of features (days in the course) changes from one year to the next, we cannot use the model trained with cluster analysis in one instance to test with data from another.

Our approach to assess cluster consistency relies on two steps. The first one, more subjective, depends on evaluating patterns and relationships among performance variables and clusters. The second one, more systematic, consists of training a classifier with labeled data from one year and test with data from another. We believe this two-step approach is robust but may not guarantee full consistency among years, which may have impacted the results.

The third and fourth limitations are of methodological nature. Multicollinearity may exist between our features, which is a possibility given that they are all forms of performance. This might have a negative impact on the robustness of our logistic regression based classifier. As such, we advise caution when interpreting this classifier's results. A workaround would be using Principal Component Analysis to convert a set of correlated variables into a smaller set of uncorrelated ones.

Another concern is that automated feature selection risks overfitting data, which might undermine model robustness and correctness. To prevent this, cross-validation is often used. However, given that we had very small samples, simply varying the number of folds in cross-validation changes the percentage of folds in which a particular attribute was selected. Thus, specifying a percentage threshold is bound to introduced error and uncertainty. Therefore, we assumed a compromise and only introduced cross-validation later, to choose an adequate classifier. To mitigate overfitting during feature selection, we used a filter method instead of a wrapper one, which is more robust to overfitting. Furthermore, testing the classifier with a sample from a different year also minimizes the problem.

Our fifth limitation concerns the fact that our approach was only tested in our learning setup. We would like to further explore its applicability to other contexts and encourage other researchers experiment and improve it in other settings. These are crucial steps towards validation.

6 Conclusion

In previous work we have identified four different types of students, characterized by distinct performance and engagement levels, behavior and gaming traits. In this paper we presented a novel experiment where we studied how we can take advantage of what differentiates different types of students, in a gamified setting, to predict their performance and behavior by midterm. To this effect we analyzed student data covering both performance measurements and gaming preferences, from one instance of the course, and used it to identify relevant features and train classifiers to test with data from another term.

Our study shows that the students' type can be predicted with up to 79% accuracy by midterm, using performance data only. However, data comprising both performance metrics and the students' player classification according to the BrainHex model was more accurate in earlier points in time, providing 66% accuracy after seven weeks and 47% even after five weeks of class.

From this study we learned a valuable lesson. Of course, in the particular case of our experiment, where student types were computed based on performance accrual, we expected the best predictors to be performance metrics. However, we learned that earlier in the course, student performance appears to be less discriminative and its predictive power can be improved by pairing it together with data that can be measured beforehand, such as their player type. We believe our study lays important groundwork for the development of adaptive gamified learning environments. These should draw on performance and gaming data to identify different student profiles in near real-time, which could be used to promptly adapt content to fit the students' needs and would be an important tool to assess student progress, for both students and instructors alike.

ACKNOWLEDGMENT

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013 and PTDC/EEI-SII/3154/2012. Gabriel Barata was supported by FCT, grant SFRH/BD/72735/2010.

REFERENCES

[1] J. Chen, "Flow in games (and everything else)," *Commun. ACM*, vol. 50, pp. 31–34, 2007.

[2] M. Csikszentmihalyi, *Flow: The psychology of optimal experience*. Harper Perennial, 1991.

[3] J. P. Gee, "What video games have to teach us about learning and literacy," *Comput. Entertain.*, vol. 1, no. 1, pp. 20–20, Oct. 2003. [Online]. Available: http://doi.acm.org/10.1145/950566.950595

[4] S. Bennett, K. Maton, and L. Kervin, "The 'digital natives' debate: A critical review of the evidence," *British Journal of Educational Technology*, vol. 39, no. 5, pp. 775–786, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8535.2007.00793.x

[5] H. F. O'Neil, R. Wainess, and E. L. Baker, "Classification of learning outcomes: evidence from the computer games literature," *Curriculum Journal*, vol. 16, no. 4, pp. 455–474, 2005. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/-09585170500384529

[6] M. Prensky, "Digital natives, digital immigrants part 1," *On the horizon*, vol. 9, no. 5, pp. 1–6, 2001. [Online]. Available: http://-www.albertomattiacci.it/docs/did/Digital_Natives_Digital_Immigrants.pdf

[7] B. Coller and D. Shernoff, "Video game-based education in mechanical engineering: A look at student engagement," *International Journal of Engineering Education*, vol. 25, no. 2, pp. 308–317, 2009. [Online]. Available: http://www.cedu.niu.edu/-~shernoff/collerShernoffIJEE.pdf

[8] P. Mcclean, B. Saini-eidukat, D. Schwert, B. Slator, and A. White, "Virtual worlds in large enrollment science classes significantly improve authentic learning," in *Proceedings of the 12th International Conference on College Teaching and Learning, Center for the Advancement of Teaching and Learning*, 2001, pp. 111–118. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=4A4C2041EBFF11311376362E286F814A?doi=10.1.1.90.4132&rep=rep1&type=pdf

[9] K. Squire, M. Barnett, J. M. Grant, and T. Higginbotham, "Electromagnetism supercharged!: learning physics with digital simulation games," in *Proceedings of the 6th international conference on Learning sciences*, ser. ICLS '04. International Society of the Learning Sciences, 2004, pp. 513–520. [Online]. Available: http://-dl.acm.org/citation.cfm?id=1149126.1149189

[10] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: defining "gamification"," in *Proceedings of the 15th International Academic MindTrek Conference Envisioning Future Media Environments*, vol. Tampere, F. ACM, 2011, pp. 9–15. [Online]. Available: http://doi.acm.org/10.1145/-2181037.2181040

[11] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon, "Gamification. using game-design elements in non-gaming contexts," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, ser. CHI EA '11. New York, NY, USA: ACM, 2011, pp. 2425–2428. [Online]. Available: http://doi.acm.org/10.1145/1979482.1979575

[12] K. Huotari and J. Hamari, "Defining gamification: a service marketing perspective," in *Proceeding of the 16th International Academic MindTrek Conference*. ACM, 2012, pp. 17–22. [Online]. Available: http://www.rolandhubscher.org/courses/hf765/readings/p17-huotari.pdf

[13] P. Brauner, A. Calero Valdez, U. Schroeder, and M. Ziefle, "Increase physical fitness and create health awareness through exergames and gamification," in *Human Factors in Computing and Informatics*, ser. Lecture Notes in Computer Science, A. Holzinger, M. Ziefle, M. Hitz, and M. Debevc, Eds. Springer Berlin Heidelberg, 2013, vol. 7946, pp. 349–362. [Online]. Available: http://-www.researchgate.net/publication/239526992_Increase_Physical_Fitness_and_Create_Health_Awareness_through_Exergames_and_Gamication._The_Role_of_Individual_Factors_Motivation_and_Acceptance/file/e0b4951c1895a30cc8.pdf

[14] Z. Fitz-Walter, P. Wyeth, D. Tjondronegoro, and B. Scott-Parker, "Driven to drive: Designing gamification for a learner logbook smartphone application," in *Proceedings of the 2013 Symposium on Gameful Design, Research, and Applications*, ser. Gamification 2013, Stratford, ON, Canada, 2013, pp. 42–49.

[15] S. Sheth, J. Bell, and G. Kaiser, "Halo (highly addictive, socially optimized) software engineering," in *Proceeding of the 1st international workshop on Games and software engineering*, ser. GAS, vol. 11, 2011, pp. 29–32. [Online]. Available: http://www.psl.cs.columbia.edu/wp-content/uploads/2011/04/halo-GAS2011.pdf

[16] G. Zichermann and C. Cunningham, *Gamification by Design: Implementing Game Mechanics in Web and Mobile Apps*. O'Reilly Media, Inc., 2011. [Online]. Available: http://books.google.pt/-books?hl=en&lr=&id=Hw9X1miVMMwC&oi=fnd&pg=PR7&dq=Gamification+by+design&ots=0pjget9qvm&sig=zz64HiGiPWw5GT7xZJtL8hd0YFs&redir_esc=y#v=onepage&q=Gamification%20by%20design&f=false

[17] KhanAcademy. (2014) https://www.khanacademy.org/ (Accessed 7 Aug 2014).

[18] Codecademy. (2014) http://www.codecademy.com/ (Accessed 7 Aug 2014).

[19] P. Denny, "The effect of virtual achievements on student engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '13. New York, NY, USA: ACM, 2013, pp. 763–772. [Online]. Available: http://doi.acm.org/10.1145/-2470654.2470763

[20] A. Domínguez, J. Saenz-de Navarrete, L. de Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz, "Gamifying learning experiences: Practical implications and outcomes," *Computers*

& Education, vol. 63, no. 0, pp. 380 – 392, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/-S0360131513000031

[21] G. Barata, S. Gama, J. Jorge, and D. Goncalves, "Engaging engineering students with gamification," in *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2013 5th International Conference on*, Sept 2013, pp. 1–8.

[22] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Improving participation and learning with gamification," in *Proceedings of the First International Conference on Gameful Design, Research, and Applications*, ser. Gamification '13. New York, NY, USA: ACM, 2013, pp. 10–17. [Online]. Available: http://doi.acm.org/10.1145/-2583008.2583010

[23] G. Barata, S. Gama, J. A. Jorge, and D. J. Gonçalves, "Relating gaming habits with student performance in a gamified learning experience," in *Proceedings of the First ACM SIGCHI Annual Symposium on Computer-human Interaction in Play*, ser. CHI PLAY '14. New York, NY, USA: ACM, 2014, pp. 17–25. [Online]. Available: http://doi.acm.org/10.1145/2658537.2658692

[24] J. Moreno, "Digital competition game to improve programming skills," *Educational Technology & Society*, vol. 15, no. 3, pp. 288–297, 2012.

[25] J. Lee, K. Luchini, B. Michael, C. Norris, and E. Soloway, "More than just fun and games: assessing the value of educational video games in the classroom," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '04. New York, NY, USA: ACM, 2004, pp. 1375–1378. [Online]. Available: http://-doi.acm.org/10.1145/985921.986068

[26] M. Kebritchi, A. Hirumi, and H. Bai, "The effects of modern math computer games on learners' math achievement and math course motivation in a public high school setting," *British Journal of Educational Technology*, vol. 38, no. 2, pp. 49–259, 2008. [Online]. Available: http://www.dynakid.com/download/DimensionM_Research_Brief.pdf

[27] J. Hamari, J. Koivisto, and H. Sarsa, "Does gamification work? – a literature review of empirical studies on gamification," in *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, Jan 2014, pp. 3025–3034.

[28] G. Richter, D. R. Raban, and S. Rafaeli, "Studying gamification: The effect of rewards and incentives on motivation," in *Gamification in Education and Business*. Springer, 2015, pp. 21–46.

[29] L. Sheldon, *The Multiplayer Classroom: Designing Coursework as a Game*. Course Technology PTR, 2011.

[30] C. Cheong, F. Cheong, and J. Filippou, "Quick quiz: A gamified approach for enhancing learning," in *Pacific Asia Conference on Information Systems*, 2013.

[31] L. Haaranen, P. Ihantola, L. Hakulinen, and A. Korhonen, "How (not) to introduce badges to online exercises," in *Proceedings of the 45th ACM technical symposium on Computer science education*. ACM, 2014, pp. 33–38.

[32] L. Hakulinen and T. Auvinen, "The effect of gamification on students with different achievement goal orientations," in *Teaching and Learning in Computing and Engineering (LaTiCE), 2014 International Conference on*. IEEE, 2014, pp. 9–16.

[33] S. Aguilar, C. Holman, and B. Fishman, "Multiple paths, same goal: Exploring the motivational pathways of two distinct game-inspired university course designs," *Games+ Learning+ Society, Madison, WI*, 2014.

[34] S. Aguilar, B. Fishman, and C. Holman, "Leveling-up: Evolving game-inspired university course design," *to appear*. [Online]. Available: http://blog.gradecraft.com/wp-content/uploads/-2014/06/Leveling-Up.pdf

[35] B. De Schutter and V. Vanden Abeele, "Gradequest—evaluating the impact of using game design techniques in an undergraduate course," in *9th International Conference on the Foundations of Digital Games*, 2014.

[36] B. De Schutter, ""the gradequest tale of scrotie mcboogerballs" evaluating the second iteration of a gameful undergraduate course." in *Meaningful Play 2014*, Michigan State University, East-Lansing, Michigan, USA, 2014.

[37] E. L. Deci, "Effects of externally mediated rewards on intrinsic motivation." *Journal of personality and Social Psychology*, vol. 18, no. 1, p. 105, 1971.

[38] R. M. Ryan and E. L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary educational psychology*, vol. 25, no. 1, pp. 54–67, 2000. [Online]. Available: http://www.selfdeterminationtheory.org/SDT/documents/-2000_RyanDeci_IntExtDefs.pdf

[39] M. R. Lepper, D. Greene, and R. E. Nisbett, "Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis." *Journal of Personality and social Psychology*, vol. 28, no. 1, p. 129, 1973.

[40] D. C. Pope, *Doing school: How we are creating a generation of stressed out, materialistic, and miseducated students*. Yale University Press, 2001.

[41] H. D. Semke, "Effects of the red pen," *Foreign language annals*, vol. 17, no. 3, pp. 195–202, 1984.

[42] E. D. Mekler, F. Brühlmann, K. Opwis, and A. N. Tuch, "Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements," in *Proceedings of the First International Conference on Gameful Design, Research, and Applications*. ACM, 2013, pp. 66–73.

[43] C. L. Diener, "Similarities and differences between over-achieving and under-achieving students," *The Personnel and Guidance Journal*, vol. 38, no. 5, pp. 396–400, 1960.

[44] W. W. Farquhar and D. A. Payne, "A classification and comparison of techniques used in selecting under-and over-achievers," *The Personnel and Guidance Journal*, vol. 42, no. 9, pp. 874–884, 1964.

[45] D. B. McCoach and D. Siegle, "Factors that differentiate underachieving gifted students from high-achieving gifted students," *Gifted Child Quarterly*, vol. 47, no. 2, pp. 144–154, 2003.

[46] D. A. Kolb *et al.*, *Experiential learning: Experience as the source of learning and development*. Prentice-Hall Englewood Cliffs, NJ, 1984, vol. 1.

[47] N. Fleming and C. Mills, "Vark: A guide to learning styles," *Retrieved November*, vol. 30, p. 2004, 2001.

[48] L. S. Larkey, "Automatic essay grading using text categorization techniques," in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 90–95.

[49] N. Pattanasri, M. Mukunoki, and M. Minoh, "Learning to estimate slide comprehension in classrooms with support vector machines," *Learning Technologies, IEEE Transactions on*, vol. 5, no. 1, pp. 52–61, 2012.

[50] B. Minaei-Bidgoli, D. A. Kashy, G. Kortmeyer, and W. F. Punch, "Predicting student performance: an application of data mining methods with an educational web-based system," in *Frontiers in Education, 2003. FIE 2003 33rd Annual*, vol. 1. IEEE, 2003, pp. T2A–13.

[51] J. Koivisto and J. Hamari, "Demographic differences in perceived benefits from gamification," *Computers in Human Behavior*, vol. 35, pp. 179–188, 2014.

[52] Moodle. (2014) Moodle. http://www.moodle.org (Accessed 7 Aug 2014).

[53] G. Barata, S. Gama, J. Jorge, and D. Gonçalves, "Gamification for smarter learning: tales from the trenches," *Smart Learning Environments*, vol. 2, no. 1, pp. 1–23, 2015. [Online]. Available: http://dx.doi.org/10.1186/s40561-015-0017-8

[54] Weka. (2014) http://www.cs.waikato.ac.nz/ml/weka/ (Accessed 31 Oct 2014).

[55] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[56] N. Sharma, A. Bajpai, and M. R. Litoriya, "Comparison the various clustering algorithms of weka tools," *facilities*, vol. 4, p. 7, 2012.

[57] L. Nacke, C. Bateman, and R. Mandryk, "Brainhex: Preliminary results from a neurobiological gamer typology survey," in *Proceedings of 10th International Conference on Entertainment Computing (ICEC'11)*, Vancouver, BC, 2011, pp. 288–293.

[58] C. Bateman and R. Boon, *21st Century Game Design (Game Development Series)*. Rockland, MA, USA: Charles River Media, Inc., 2005.

[59] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, 1999.

[60] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[61] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

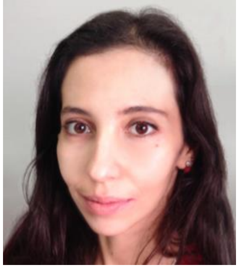
[62] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[63] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[64] T. Mehdi, N. Bashardoost, and M. Ahmadi, "Kernel smoothing for roc curve and estimation for thyroid stimulating hormone," *International Journal of Public Health Research*, pp. 239–242, 2011.

[65] N. V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 853–867.

Gabriel Barata is a researcher at the Visualization and Multimodal Interfaces Group of INESC-ID. He holds a PhD in Information System & Computer Engineering from Instituto Superior Técnico (IST), University of Lisbon (UL), Portugal. Gabriel has been studying the impact of gamification on college education and how it can be used to identify different student profiles and help developing adaptive learnings environments. His research interests include human-computer interaction, personal information management and gamification.



Sandra Gama is a researcher at the Visualization and Multimodal Interfaces Group, INESC-ID, and a Professor of Computer Science at Instituto Superior Técnico, University of Lisbon, Portugal. Her research falls within the scope of Information Visualization, Multimodal User Interfaces and Human-Computer Interaction.

Joaquim Jorge received his PhD from Rensselaer Polytechnic Institute in 1995, coordinates the VIMMI research group at INESC-ID and is Full Professor of Computer Graphics and Multimedia at *Técnico*, *Universidade de Lisboa*. He is Editor-in-Chief of the Computers and Graphics Journal (Elsevier), a Fellow of the Eurographics Association and Senior Member of ACM and IEEE, serves on the ACM Europe Council and Chairs the ACM/SIGGGRAPH Specialized Conferences Committee. He organized 35+ international scientific events, is Eurographics 2016 papers cochair, served on 180+ program committees and (co)authored 250+ publications in international refereed venues. His research interests include multimodal user interfaces, 3D Visualization and advanced learning techniques.

Daniel Gonçalves is a researcher at the Visualization and Multimodal Interfaces Group of INESC-ID and Professor of Computer Science at Instituto Superior Técnico (IST/UL), Portugal. His research interests include the areas of Information Visualization, Personal Information Management, Human-Computer Interaction and Accessibility. He is a member of ACM and the Portuguese Computer Graphics Group (the national Eurographics chapter).