Education and Information Technologies (2023) 28:5209–5240
https://doi.org/10.1007/s10639-022-11403-y

The use of video clickstream data to predict university students' test performance: A comprehensive educational data mining approach

Ozan Raşit Yürüm[1] · Tuğba Taşkaya-Temizel[2] · Soner Yıldırım[3]

Received: 21 March 2022 / Accepted: 12 October 2022 / Published online: 29 October 2022

Abstract

Video clickstream behaviors such as pause, forward, and backward offer great potential for educational data mining and learning analytics since students exhibit a significant amount of these behaviors in online courses. The purpose of this study is to investigate the predictive relationship between video clickstream behaviors and students' test performance with two consecutive experiments. The first experiment was performed as an exploratory study with 22 university students using a single test performance measure and basic statistical techniques. The second experiment was performed as a conclusive study with 16 students using repeated measures and comprehensive data mining techniques. The findings show that a positive correlation exists between the total number of clicks and students' test performance. Those students who performed a high number of clicks, slow backward speed or doing backwards or pauses achieved better test performance than those who performed a lower number of clicks, or who used fast-backward or fast-forward. In addition, students' test performance could be predicted using video clickstream data with a good level of accuracy (Root Mean Squared Error Percentage (%RMSE) ranged between 15 and 20). Furthermore, the mean of backward speed, number of pauses, and number/percentage of backwards were found to be the most important indicators in predicting students' test performance. These findings may help educators or researchers identify students who are at risk of failure. Finally, the study provides design suggestions based on the findings for the preparation of video-based lectures.

Keywords Educational data mining · Learning analytics · Performance prediction · University students · Video clickstream interactions

✉ Ozan Raşit Yürüm
ozanyurum@iyte.edu.tr

Extended author information available on the last page of the article

1 Introduction

Video-based learning has become widely popular and accessible due to the rapid developments in Massive Open Online Courses (MOOCs) and video-sharing platforms such as YouTube and Vimeo (Ilioudi et al., 2013). Educational films, educational television programs, and video cassettes also historically contributed to the development of the video-based learning (Kleftodimos, 2016; Yousef et al., 2014). The availability of increased internet network bandwidth and the variety of digital devices now in play have facilitated widespread access to video lectures for the masses (Ronchetti, 2013). In addition, as a consequence of the COVID-19 pandemic, many educational institutions have adopted video-based lecturing as a core element of their distance education model (El Aouifi et al., 2021). The field of video-based learning analytics has emerged to measure the effectiveness of educational videos since video clickstream behaviors can be used together with analytical approaches to improve learning (Giannakos et al., 2013a).

Analysis of video viewing behaviors based on clickstream interactions such as pause, forward, and backward can reveal insights into students' preferences and learning styles (De Boer et al., 2011; Dissanayake et al., 2018). Understanding these behaviors can aid the design of learning environments (Li et al., 2015), and for implementing timely interventions aimed at improving students' learning performance (Kleftodimos & Evangelidis, 2014; Lu et al., 2018).

A limited number of studies have investigated video clickstream behaviors on video-based learning platforms to predict students' learning or test performance (Atapattu & Falkner, 2018). A few studies have tracked the video clickstream behaviors of learners without elaborating on the relationship between video clickstream behaviors and the students' academic success. In addition, the existing studies generally employ a single measure to analyze video clickstream behaviors. The current study investigates the correlations between video clickstream behaviors and students' test performance based on repeated measures. It also aims to predict grades and to reveal significant predictors of students' test performance by using data mining techniques that have yet to be explored in detail in the context of educational data mining; a relatively new field of study that aims to improve learning through the mining for and analysis of data from educational settings. The output of the current study is suggested to be of value in the designing of video lectures as well as for intervention-related video-based studies.

In addition, the proposed approach may also be applied to e-learning systems since it provides input for real-time interventions during video lectures. As the number of courses in such platforms increases at a fast pace, there is usually no data or only a limited amount available from these new courses. However, the proposed method shows that behavioral data from previous related videos may be used to infer users' performance on such platforms; making it possible for necessary timely interventions to be applied during the teaching and learning process. When there is sufficient data for users and courses, the proposed method is expected to perform well in terms of predicting students' test performance. The results of the current study are also expected to be of significant value for researchers interested in learning analytics and educational data mining, as they may prove useful as a guide for the application and evaluation of these techniques within the education domain.

2 Literature review

The literature is presented in two complementary subsections within the scope of the study: (1) Video-Based Learning Analytics and (2) Prediction of Students' Performance. The first addresses the advantages of using analytics in video-based learning and the studies based on users' watching behaviors. The second subsection presents the previous studies about learning or test performance prediction based on video clickstream behaviors.

2.1 Video-based learning analytics

Video-based learning analytics, or video analytics for short, can be defined as "the collection, measurement, and analysis of such data for the purposes of understanding how learners use videos" (Mirriahi & Vigentini, 2017, p. 251). It emerged as a subfield of learning analytics; thereby, video-based learning analytics targets the enhancement of learning using educational data obtained from platforms that offer video lectures (Giannakos et al., 2016).

The Workshop on Analytics on Video-Based Learning (WAVe; Giannakos et al., 2013b), which was organized as part of the 3rd Conference on Learning Analytics and Knowledge (LAK; Suthers et al., 2013), pioneered the use of video analytics in learning since it was the initial formal event held within the field. Brooks et al. (2013) applied visualization techniques that mostly showed rewatching behaviors or video watching time to reveal students' watching behaviors. Chorianopoulos and Giannakos (2013) defined a novel video analytics tool that captured learners' interaction with video lectures. Moreover, Ronchetti (2013) explained potential improvements to the subject of video analytics, and highlighted how meaningful data may be extracted from the videos.

In 2014, Giannakos et al. introduced a further improved tool named SocialSkip, and discussed how video analytics could be applied in the field of learning. In that same year, other significant studies that contributed to the literature in this area were also presented at the Workshop on Modeling Large Scale Social Interaction in Massive Open Online Courses held at the Conference on Empirical Methods in Natural Language Processing (EMNLP; Rose & Siemens, 2014). The presented studies generally aimed their focus on predicting student drop-out in MOOCs, with some applying different types of machine learning algorithms to predict drop-out based on video clickstream behaviors. Sinha et al. (2014a) identified a group of behavioral actions leading to high or low levels of information processing. In a separate study, Sinha et al. (2014b) employed both video lecture clickstream behavior data and discussion forum activities to predict drop-out to a moderate level. Kloft et al. (2014) employed various data resources in addition to videos, and stated that whilst the prediction accuracy of drop-out can be seen as low at the beginning, it can increase over time.

These studies present good examples of how video clickstream behaviors can yield significant results, and especially for MOOC designers. In addition to these studies, Kleftodimos and Evangelidis (2014) examined video viewing behaviors in order to cluster learner engagement and video popularity. They identified two clusters based on video popularity and seven for learner engagement. However, they also suggested the need for further investigation on the factors that can affect these behaviors.

Giannakos et al. (2015) focused on the number of views per second of video lectures, and revealed a relationship between repeated views and the level of cognition/thinking required for a particular video segment. Li et al. (2015) also investigated the relationship between video interaction patterns and certain variables deemed important such as perceived video difficulty and learning performance. Their findings revealed a relationship between some of these patterns or clusters and the aforementioned variables.

Kleftodimos (2016) demonstrated that a click-based learning environment may be used to track both video clickstream behaviors and navigational behaviors for learning analytics, and went on to cluster these behaviors and explain each cluster. Then, Dimitrova et al. (2017) took video analytics that one step further and attempted to develop nudges from video comments in order to increase active watching times.

Lau et al. (2018) used mainly view counts and mean percentages as a way to evaluate video lectures, and from that showed how students' retention decreased over time. Atapattu and Falkner (2018) used video clickstream behaviors in their research, and revealed a positive correlation between several discourse features and video interactions. In addition, Akçapınar and Bayazıt (2018) investigated video viewing behaviors and concluded that surface learners and deep learners performed different types of video interactions.

Yoon et al. (2021) also applied video analytics as a means to investigate behavioral patterns and learner clusters from watching video lectures. Their study revealed four patterns that may prove useful in determining active versus passive learners. Furthermore, Yürüm et al. (2022) also used video clickstream behaviors to develop an intervention framework that could be utilized to transform linear videos into interactive videos.

In summary, video-based analytics have been employed for descriptive, predictive, and prescriptive purposes. Descriptive studies use historical video viewing behaviors in order to understand student behaviors, with studies generally focused on finding behavioral patterns and attempting to explain them. Predictive studies use video analytics to predict a set of learning variables such as video popularity and drop-out, whilst prescriptive studies utilize video analytics to develop interventions aimed at increasing the efficiency of video lecturing.

Considering these varied prior research, the following benefits to video-based learning analytics stand out in their significance:

- To better understand learner behaviors and to design video lectures aimed at increasing these behaviors.
- To predict learning variables in order to identify students who are at risk of failure/drop-out.
- To intervene in a timely manner so as to increase student success and retention.

In order to realize these benefits, studies have utilized general properties such as video popularity, watching time, and numbers of views or clickstream interactions such as pause, forward, and backward. Atapattu and Falkner (2018) classified these video-watching behaviors as either implicit factors (e.g., pause, backward, etc.) or explicit factors (e.g., views, etc.), but specifically they emphasized there being a research gap in the literature regarding the use of implicit factors. The published studies that have focused on clickstream interactions or implicit factors have concentrated either on transitions or sequences between click types in order to identify similar user behaviors or the number of clickstream interactions.

Various video clickstream interactions have been employed in studies based on students' video-watching behaviors. These video clickstream interactions have also been referred to as "events" in some studies. Whilst the existing studies generally include some of these behaviors, it was also discovered that play and pause are the most widely used interactions in research studies as well as backward and forward have also been utilized in a significant number of studies (Seidel, 2017). Notably, "rewind" is also used interchangeably with backward, whilst "seek" is used to represent both backward and forward in certain studies.

2.2 Prediction of students' performance

Developing prediction models or determining predictors of learning performance is shown to help identify students who are at risk of academic failure or drop-out (Mubarak et al., 2021). These predictions can provide insights regarding the timing of interventions (Lu et al., 2018). Numerous studies have focused on predicting students' learning or test performance (Brinton et al., 2015; Guo et al., 2015; Hussain et al., 2022; Soni et al., 2018), with most having used navigational or demographic data (Gardner & Brooks, 2018; Shahiri et al., 2015). However, only a limited number of studies have focused on the video clickstream behaviors of students in order to predict their performance.

Ullrich et al. (2013) aimed to predict a set of variables with a k-Nearest Neighbors (kNN) classifier by using students' viewing patterns that were determined based on video clickstream behaviors such as backward and forward. Their study showed that some predictions, including whether or not a student would pass, are possible, even if the prediction accuracy is at just a moderate level (F1 score = 0.63). In their study, Ullrich et al. (2013) used certain default settings for their predictions and emphasized the importance of detailed analysis, including parameter tuning to obtaining better models that utilized machine learning approaches.

Brinton et al. (2015) studied the relationship between sequential clickstream interaction patterns and Correct on First Attempt (CFA) in a quiz. They found that certain behavioral patterns have a positive relationship with CFA, and that three prediction models constructed based on transitions or positions of the interactions proved to be significantly more successful than the baseline model. Nevertheless, the accuracy of the models ranged from 0.64 to 0.66, and with F1 scores between 0.69 and 0.74, but the study did not focus on the individual predictive relationship between learning performance and video clickstream behaviors.

Li et al. (2016) used additional features derived from personal, exercise, and quiz information together in addition to video clickstream behaviors in order to predict students' grades in MOOCs. They applied regression and neural network models to deliver predictions with a high degree of accuracy. However, a significant amount of information is required in order to put the model into practice. In addition, their results may be said to have been expected, since the students' past performance information was used as input in building the study's prediction models.

Yang et al. (2017) also employed recurrent neural networks to predict students' test performance, and used past quiz performance data and clickstream data together as input to the prediction of average CFA quiz scores. However, aggregating numerous output measurements into one single average measure may be said to go somewhat against the nature of employing repeated measures.

Lu et al. (2018) investigated several features such as out-of-class practice behaviors, homework, and quiz scores, as well as video clickstream behaviors to predict students' learning performance for a blended course on calculus. The authors revealed seven critical factors, including the number of backward clicks per week and the number of plays per week in their study.

Yu et al. (2019) used seven behavioral patterns based on the sequences of viewing behaviors to predict students' pass/fail scores with kNN, Support Vector Machines (SVM), and Artificial Neural Network (ANN) algorithms, with ANN providing the highest rate of prediction accuracy.

The study published by Hasan et al. (2020) aimed to predict higher education students' performance by utilizing various features such as student academic information, students' activities, and students' video interactions, and then compared their predictive performance using various data mining algorithms. The Random Forest algorithm was shown to provide the best results. However, the individual impact of video clickstream behaviors on students' test performance could not be inferred from the study's results since the students' previous coursework scores and exam scores were also included as input into the prediction models. Mbouzao et al. (2020) also focused on video interaction features in their study, and used time-related metrics to predict whether students would likely pass or fail; showing that these metrics could be employed to predict students' performance. Van Goidsenhoven et al. (2020) utilized a number of features, some of which are related to navigation, forum, and problem to predict student success in a blended learning environment. However, a considerable amount of data is needed in order to properly apply the prediction model.

El Aouifi et al. (2021) focused on video viewing sequences and reportedly predicted learning performance with limited accuracy (65.07%) using educational data mining techniques; whereas Mubarak et al. (2021) used deep neural network (LSTM), but revealed no predictive relationship between each of the analyzed video clickstream behaviors and learning performance separately.

As can be understood from the current literature, many different algorithms have been used in various attempts to predict learning performance. While some have focused only on video clickstream behaviors, others also used additional features. However, as yet, it is not possible to infer which video clickstream behavior is the more critical when compared to all the others. Moreover, many studies have used either similar features or aggregated features obtained from several videos, but disregarded the notion that students may of course exhibit different behaviors for different videos. The existing studies also do not provide any detailed or systematic explanation as to the approaches employed in their analyses, which inherently impairs the potential reproducibility of the results.

The current study, however, differs from the existing literature in that it aims to show the significant features of video clickstream behaviors and their relationships to predict students' test performance. These behaviors are defined in numerous dimensions, such as the number of clicks, the relative percentage of clicks, and forward and backward clicking speeds. Repeated measures analysis is used to model the repeated measures obtained from the students, which successfully considers both within-subject and between-subject factors. Finally, the XGBoost model is applied in predicting scores using a comprehensive evaluation strategy, and significant features are also shown. This study explains the applied analytical approaches with step-by-step detail.

3 Research method

The current study includes two experiments based on the one-group posttest-only design (also termed "one-shot case study design" by Fraenkel et al. (2012, p. 269)). The first one takes a single measure and aims to reveal whether or not a correlation exists between the video clickstream behaviors and test performance. Feedback from the first experiment is then used to design the second experiment as a repeated measure study. The first experiment also showed potential improvement points for the development of the second experiment. Table 1 defines the steps performed during this research:

3.1 Research questions

The research questions of the study are based on video clickstream behaviors and students' test performance. The students' test performance data can be obtained through post-tests, whilst video clickstream behaviors can be tracked up until the post-tests are administered. One single test group is sufficient since the study is focused on predictive relationships, but does not go on to compare the impact of any intervention on different groups; hence, the one-group posttest-only design was selected for the current study. Applying an exploratory experiment strengthened the second experiment in order to reach conclusions to answer the study's research questions. The second experiment provides repeated measures so as to answer the research questions using comprehensive data mining techniques (see Table 1).

The following two research questions were formed to guide the study:

RQ1 *Are there any correlations between students' video clickstream behaviors and their test performance?*

Table 1 Research steps

Study	Method	Type	Analysis techniques	Steps
1. First Experiment	One-group posttest-only design	Exploratory	Basic statistical techniques	1. Student watches three video lectures via the learning management system (LMS). 2. Student takes a midterm exam.
2. Second Experiment	One-group posttest-only design with repeated measures	Conclusive	Comprehensive data mining techniques	1. Student watches the first video lecture via an online platform. 2. Student takes the first video quiz. 3. Student watches the second video lecture via an online platform. 4. Student takes the second video quiz. 5. Student watches the third video lecture via an online platform. 6. Student takes the third video quiz.

RQ2 *Which video clickstream behaviors and to what extent can these behaviors predict students' test performance?*

For each research question, several sub-questions were also developed. For the first research question (RQ1), the number of clicks and the speed of clicks form the main focus of the study. Therefore, the following sub-questions further define the first research question:

RQ1.1 *Is there a significant correlation between the number of clicks (total, pause, backward, and forward) and the students' test performance?*

RQ1.2 *Is there a significant correlation between the speed of clicks (mean backward speed and mean forward speed) and the students' test performance?*

The second research question aims to establish which of the learners' video clickstream behaviors better predict their learning performance. Two different models were attempted. The first, XGBoost, a well-known machine learning model, was chosen for modeling since it has been shown in the literature to offer superior performance in many data mining tasks (Shi et al., 2019; Tadesse et al., 2018). The second model, the Robust Linear Mixed Model, is a well-known statistical method that can model repeated measures and also cope with contamination or outliers on any level. Baseline models were also constructed for comparison purposes. Therefore, this second research question includes the following sub-questions:

RQ2.1 *Does the XGBoost algorithm predict students' test performance better than the baseline models?*

RQ2.2 *Does the Robust Linear Mixed Model predict students' test performance better than the baseline models?*

3.2 First experiment

The first experiment aimed to measure whether or not a correlation exists between students' video clickstream behaviors and their test performance, and to help design the second experiment.

3.2.1 Participants

The convenience sampling strategy was applied in this study because the participants were easily accessible to the researchers. The first experiment was performed with 28 students registered to the Principles and Methods of Instruction (CEIT 216) course offered at the Department of Computer Education and Instructional Technology. A total of 22 participants watched the video lectures.

3.2.2 Data collection instruments

Two data types were collected in the study, (1) Students' video clickstream behavior data and (2) Students' midterm grade data. The clickstream behaviors of the students were collected via an LMS over a 3-week period where a video analytics tool was integrated into the LMS. The test performance scores of the students were obtained from the midterm exam which the students took within a face-to-face classroom environment.

The metrics collected via the LMS included the video clickstream behaviors based on the participant students' clickstream interactions for *Play, Pause, Forward,* and *Backward*. The midterm exam consisted of 40 questions, of which 35 were multiple-choice, and five were true/false type questions. The midterm exam, which was conducted as a closed-book exam, included questions about the content that was provided in the course video lectures.

The video lectures were on the subject of famous educational psychologists, and the same producer developed each video. The videos in the first experiment were: (1) "Piaget's Developmental Theory: An Overview" (27:06 min), (2) "Play: A Vygotskian Approach" (26:13 min), and (3) "Maria Montessori: Her Life and Legacy" (35:10 min).

3.2.3 Procedure

In total, 22 students watched the videos linearly over a 3-week period in their own time. The videos were embedded into the LMS, with the students' basic clickstream behaviors tracked by the plug-in application called "SocialSkip" (Chorianopoulos et al., 2011). Two weeks after the video-watching period, the students each sat a scheduled midterm exam that was significantly weighted in terms of the overall course grade. The students' midterm exam scores were graded out of 100. Table 2 presents the number of interactions (video clickstream behaviors) that the students performed for each video. There were a total of 2,192 clicks performed by all the participants while watching all three videos of the first experiment. The total number of clicks for play, pause, backward, and forward were 930, 862, 264, and 136, respectively.

Table 2 Number of interactions per interaction type for each video used in the first experiment

Video	Play	Pause	Backward	Forward	Total
1	370	348	125	63	906
2	376	338	94	56	864
3	184	176	45	17	422
Total	930	862	264	136	2,192

3.2.4 First experiment results

A positive relationship was found between the total number of interactions (i.e., all interaction types) that the students performed and the test performance (*the correlation coefficient* (*r*) $= 0.59$, $p < 0.01$). Figure 1 shows the correlations between interaction numbers per interaction type and test performance. In particular, both the number of backward click interactions ($r = 0.65$, $p < 0.01$) and the number of pauses ($r = 0.58$, $p < 0.01$) were found to be highly correlated with the students' test performance. The number of forward clicks, however, was not found to be significantly correlated with the performance.

This study was conducted with a single measure (one midterm score). There may be various factors that could affect the students' midterm exam scores. Therefore, a second experiment was designed in which the students were set a quiz immediately following each video to better observe the relationships between the features.

The first experiment also showed that a dedicated video analytics tool, having more metrics and tracking features, might be better in revealing students' video watching behaviors. Therefore, additional interaction types such as "full screen on/off" or "mute/unmute" were planned for the second experiment.

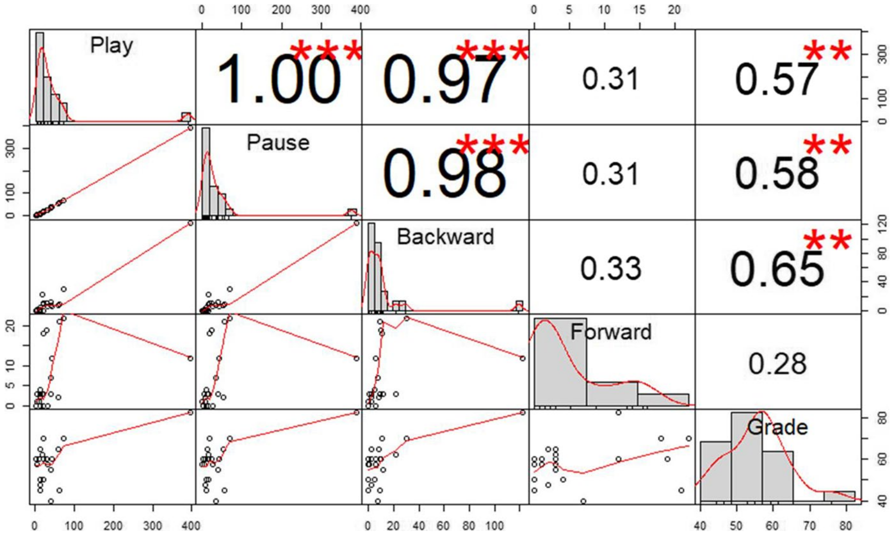

Fig. 1 Pearson correlation results between the interaction numbers per interaction type (all video lectures) and students' test performance (midterm grade)

3.3 Second experiment

The study's second experiment adopted a repeated measure approach to analyze the results.

3.3.1 Participants

The convenience sampling method was also used in the second experiment. The participants were students registered to a course on Instructional Principles and Methods (EDS212) that was given during the 2019/2020 Fall semester, and who were neither participants in the study's first experiment (the content was the same as the course content in the first experiment.), nor had they taken the course previously. There were initially 22 registered students; however, two later dropped the course. Sixteen of the remaining students watched more than one video lecture. Table 3 presents the number of interactions for each participant based on each video. For example, the first participant (P01) performed 89 interactions (including 35 plays, 34 pauses, one forward, and 12 backwards, plus others such as volume up, full screen on, and mute) in the first video lecture, 29 interactions in the second video lecture, and 118 interactions in the third video lecture. The total number of interactions for the first participant were therefore 236.

Table 3 Number of interactions for each participant (with rows showing interaction statistics for each individual)

	First video lecture					Second video lecture					Third video lecture					Total
	NPl	NP	NF	NB	T	NPl	NP	NF	NB	T	NPl	NP	NF	NB	T	
P01	35	34	1	12	89	9	7	0	5	29	53	52	0	6	118	236
P02	-					3	2	0	1	13	2	1	0	0	10	23
P03	2	1	0	0	5	-					2	1	5	0	15	20
P04	4	3	0	1	11	6	6	3	4	31	24	23	0	8	65	107
P05	-					11	11	5	15	50	16	15	15	17	68	118
P06	22	17	32	30	127	20	18	17	41	104	24	23	6	26	86	317
P07	3	0	0	0	10	6	5	10	0	30	1	0	0	0	2	42
P08	3	0	0	0	11	4	0	0	0	11	33	31	3	22	100	122
P09	16	16	2	11	48	8	7	1	3	25	-					73
P10	-					3	2	7	0	16	2	2	8	1	15	31
P11	67	67	6	13	160	64	63	17	46	198	3	2	0	0	8	366
P12	49	48	1	9	117	-					7	6	15	1	32	149
P13	9	6	3	20	46	3	2	0	2	9	6	4	0	3	21	76
P14	9	8	7	9	38	9	8	0	3	23	14	13	17	11	58	119
P15	1	1	0	1	5	18	18	7	12	58	11	11	3	7	40	103
P16	4	3	0	0	11	4	4	0	1	11	8	7	1	0	28	50
Total	228	206	52	106	692	151	138	55	117	529	186	173	50	84	573	1,794

NPl = # of Plays, NP = # of Pauses, NB = # of Backwards, NF = # of Forwards, T = Total Interactions (including other types)

3.3.2 Data collection instruments and procedure

The clickstream behaviors of the participants were collected via an online platform including the additional metrics of “volume up/down,” “mute/unmute,” and “full screen on/off.” The students’ performances were measured with a quiz after they had watched each video lecture. A total of three video lectures from the same producer were presented: (1) “Maria Montessori: Her Life and Legacy” (35:10 min), (2) “Erik H. Erikson: A Life’s Work” (30:42 min), and (3) “John Dewey: His Life and Work” (29:55 min). It was observed that play, pause, backward, and forward were the most performed interactions in the study. However, the students did not perform a noticeable number of newly added interactions during their watching of the three video lectures. In addition, since the number of pauses and the number of plays were highly correlated, only the basic interactions of pause, backward, and forward were considered (except for calculating the total number of clicks). The features were derived from these three interaction types during the analysis phase of the study.

The video quizzes, as related to the content of each specific video lecture, included multiple-choice and true/false type questions. The difficulty level of all three quizzes matched; however, they consisted of different questions at various levels compared to the first experiment. The number of questions varied from eight to nine according to the video lecture, and each quiz was graded out of 100 points. Table 4 shows the number of interactions that the students performed for each video, with Video 1 having received a higher number of interactions than the other two.

4 Results

The results of this study are presented in two subsections: (1) Relationship between Video Clickstream Behaviors and Test Performance, and (2) Predicting Test Performance with XGBoost and Robust Linear Mixed Models. The first shows the relationship results for repeated measures, whilst the second shows the predictors of test performance and to what extent students’ test performance can be predicted based on video clickstream behaviors using the two different models.

Table 4 Number of interactions in the second experiment

Video	Play	Pause	Forward	Backward	Total
1	228	206	52	106	692
2	151	138	55	117	529
3	186	173	50	84	573
Total	565	517	157	307	1,794

4.1 Relationship between video clickstream behaviors and test performance

The first research question aimed at investigating the relationship between the students' video clickstream behaviors and their test performance. The research sub-questions measured the correlation between the total number of interactions, number of pauses, number of backward clicks, number of forward clicks, the mean backward speed, the mean forward speed, and their test performance. Repeated measures correlation analysis was employed, which is a statistical method used to evaluate the general within-individual relationship between paired measures that are being measured on two or more occasions (Bakdash & Marusich, 2017). The Repeated Measures Correlation (rmcorr) package was used in R.

The *speed* for *backward clicks BS,* and the *speed* for *forward clicks FS* for each interaction were calculated by dividing the amount of skipped time by each video's duration:

$$BS_{i,j} = BT_{i,j}/VT_{i,j}$$
$$FS_{i,j} = FT_{i,j}/VT_{i,j}$$

where i = related user id, j = related video lecture id, BT = mean skipping time for backward clicks, FT = mean skipping time for forward clicks, and VT = video duration.

Since all 16 students did not watch all three video lectures, there were a total of 42 observations subjected to analysis, rather than 48 (16 students × 3 videos). In total, 29 observations included backward-click interactions, whilst 25 observations had forward-click interactions, after removing a few backward and forward click interactions that did not reflect the student's interaction pattern in terms of their usual speed. Therefore, the number of observations analyzed varied according to the research sub-question.

For *RQ1.1*, the results revealed

- a significant positive correlation between the total number of clicks and test performance ($n = 42$, $r = 0.66$, $p = 0.00$).
- a significant positive correlation between the number of pauses and test performance ($n = 42$, $r = 0.62$, $p = 0.00$).
- a significant positive correlation between the number of backward clicks and test performance ($n = 42$, $r = 0.48$, $p = 0.01$).
- no significant correlation between the number of forward clicks and test performance ($n = 42$, $r = 0.32$, $p = 0.10$).

For *RQ1.2*, the results revealed

- a significant negative correlation between the mean backward speed and test performance ($n = 29$, $r =$ -0.65, $p < 0.01$).
- no significant correlation between the mean forward speed and test performance ($n = 22$, $r =$ -0.28, $p > 0.05$).

4.2 Test performance prediction with XGBoost and robust linear mixed models

In this section, the XGBoost (Chen & Guestrin, 2016) and Robust Linear Mixed Model (Koller, 2016) algorithms were employed. XGBoost was preferred since it is known to be scalable and effective using gradient boosting trees, and works well with small-sized datasets (Chen et al., 2015). Robust LMM was preferred as it makes no assumptions, with the exception that the model parameters are estimable and considered suitable for repeated measures (Koller, 2016). Especially, smaller-sized samples have the potential to include outliers or non-Gaussian distributions (Demidenko, 2013); hence, Robust LMM can be used for non-normal distributed residuals when the assumptions for LMM are not met (Schielzeth et al., 2020).

As an observation may belong to any one of the 16 participant students, multiple observations for each student may not be able to be treated as independent and residuals may be considered non-normal. Hence, a Robust Linear Mixed Method (RLMM) model was selected to be used as a linear additive model for multiple levels by groups, where the random effect within group is additive to the fixed effect between groups:

$$Y_i = X_i\beta + Z_ib_i + \varepsilon_i$$

where Y is the vector of the response variable (quiz grade), X is the fixed effect (video interaction features), β is the vector of the fixed effects coefficients, Z is the random effects (each user), b is the vector of random effects coefficients, i is the related group, and ε is the error for random effects (Fox, 2002).

Eight features related to the video clickstream behaviors of the students were used as input to predict their test performance scores. These are each related with the number, percentage, and speed based on the three click types: *Number of Pauses, Number of Forwards, Number of Backwards, Mean Forward Speed, Mean Backward Speed, Percentage of Pauses, Percentage of Backwards,* and *Percentage of Forwards.*

In addition, the features detailed in Section 4.1, Percentage of Pauses (*PP*), Percentage of Forwards (*PB*), and Percentage of Backwards (*PB*), were each calculated by dividing the number of interactions for the related click type by the total number of interactions for these three types as follows:

$$\begin{aligned} PP_{i,j} &= NP_{i,j}/(NP_{i,j} + NB_{i,j} + NF_{i,j}) \\ PB_{i,j} &= NB_{i,j}/(NP_{i,j} + NB_{i,j} + NF_{i,j}) \\ PF_{i,j} &= NF_{i,j}/(NP_{i,j} + NB_{i,j} + NF_{i,j}) \end{aligned}$$

where i = related user id, j = related video lecture id, PP = % Pauses, PB = % Backwards, PF = % Forwards, NP = # Pauses, NB = # Backwards, and NF = # Forwards

The following steps were performed for each model:

1. Feature Selection
2. Comparison with a Baseline Model (Nested 5-fold Stratified Cross-Validation)
3. Comparison with a Baseline Model (One Person Leave Out)
4. Comparison with a Baseline Model (One Video Leave Out)

Table 5 Hyperparameter tuning (parameters and values) for XGBoost used for a grid search

Parameter	Values
learning_rate	0.01, 0.1
max_depth	3, 5, 7
min_child_weight	1, 2, 3, 4, 5
subsample	0.5, 0.7
colsample_bytree	0.5, 0.7
n_estimators	400, 500, 600
objective	Mean Squared Error

Table 6 Importance of features obtained with XGBoost

Feature	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Mean	Rank
Number of Pauses	0.145	0.125	0.176	0.174	0.143	0.153	1
Number of Backwards	0.134	0.108	0.133	0.181	0.195	0.150	2
Mean Backward Speed	0.153	0.170	0.166	0.132	0.104	0.145	3
Number of Forwards	0.144	0.086	0.121	0.138	0.131	0.124	4
Percentage of Pauses	0.109	0.207	0.097	0.101	0.097	0.122	5
Percentage of Backwards	0.108	0.106	0.130	0.112	0.118	0.115	6
Mean Forward Speed	0.142	0.102	0.119	0.083	0.109	0.111	7
Percentage of Forwards	0.066	0.095	0.059	0.079	0.101	0.080	8

4.2.1 Feature selection

Three steps were followed to select the input features for model building: (1) parameter tuning (not necessary for Robust LMM), (2) feature ranking, and (3) determining the number of features respectively for both algorithms.

Step 1: Parameter Tuning for XGBoost

To find the best parameters, the dataset was split into a training dataset and a testing dataset using stratified random sampling with five folds. Stratified means the distribution of test performance scores in the original dataset was preserved in the samples. The parameters giving the best scores on the validation dataset were selected. Table 5 shows both the parameters and the values used for hyperparameter tuning of the XGBoost algorithm. This configuration was used for all the experiments in the study.

Step 2: Feature Ranking with XGBoost

Using the XGBoost algorithm, the features were ranked according to the mean feature importance rate in five iterations.

Step 3: Determining the Number of Features for XGBoost

Table 7 Comparison of three features and eight features with XGBoost

No	Model	*N* (Observations)	RMSE	MAE	SD	Z	*p*
1	XGBoost with 3 features	42	14.56	11.11	9.53	605	0.055
2	XGBoost with 8 features	42	16.38	12.75	10.40		

Table 8 Importance of the selected features obtained with XGBoost

Feature	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Mean	Rank
Number of Pauses	0.392	0.355	0.419	0.379	0.336	0.376	1
Mean Backward Speed	0.345	0.376	0.332	0.302	0.265	0.324	2
Number of Backwards	0.263	0.269	0.249	0.319	0.399	0.300	3

* $\mathrm{VIF}_1 = 1.53$, $\mathrm{VIF}_2 = 1.03$, $\mathrm{VIF}_3 = 1.53$

Root Mean Square Error (RMSE) scores were calculated for each feature set iteratively to determine the optimum number of features. The features were added one by one according to their rank. The top three features were selected as they had the lowest RMSE and were differentiated from the others in terms of their feature importance (see Mean values in Table 6). In addition, Wilcoxon Signed Rank Test, a nonparametric test for comparing two-related groups (Field, 2018), was performed as a significance test. Table 7 shows that using the first three features (RMSE = 14.56, MAE = 11.11, *Standard Deviation (SD)* = 9.53) gave better results (less Mean Absolute Error) than using all eight features (RMSE = 16.38, MAE = 12.75, *SD* = 10.40, the *Standardized Test Statistic (Z)* = 605, *p* = 0.055) even though the results were not found to be significant. This shows that adding more than three features did not improve the model and justified the feature selection.

As a result, the top three features were selected and reordered among themselves based on their feature importance (see Table 8):

Multicollinearity was also checked for each feature with Variance Inflation Factor (VIF). If VIF is less than 5, it indicates no multicollinearity (Akinwande et al., 2015).

Step 1: Feature Ranking with Robust LMM

Using the Robust LMM, the features were ranked according to their significance value in the model. Table 9 shows the significance of each feature according to its rank.

Step 2: Determining the Number of Features for Robust LMM

One variable can be repressed by another due to interaction between the variables. Therefore, backward feature selection was used to select the variables for

Table 9 Significance of the features with Robust LMM

Rank	*Variable*	*Estimate*	*Standard Error (SE)*	*t*	*p*
0	Intercept	72.57	7.55	9.61	0.00
1	Mean Backward Speed	-161.51	58.99	-2.74	0.01
2	Percentage of Backwards	37.59	18.51	2.03	0.05
3	Number of Pauses	0.26	0.21	1.28	0.21
4	Percentage of Pauses	7.63	10.08	0.76	0.45
5	Percentage of Forwards	18.99	27.72	0.69	0.50
6	Mean Forward Speed	0.11	0.79	0.14	0.66
7	Number of Forwards	-53.15	119.40	-0.45	0.89
8	Number of Backwards	-0.07	0.53	-0.13	0.90

the final model. Features were removed from the model according to their rank on a one-by-one basis until RMSE increased, or a feature exceeded the *p*-value threshold. As a result, the first three features were selected since both RMSE was the lowest and the *p*-values were within the threshold for the three features. Wilcoxon Signed Rank Test showed that using three features (RMSE = 15.11, MAE = 11.27, $SD = 10.19$) provided better results (less MAE) than using all eight features (RMSE = 16.40, MAE = 12.66, $SD = 10.54$, $Z = 462$, $p = 0.185$) (see Table 10). There was no need to utilize all eight features since adding the additional features would not improve the model.

Since the overall goal was to predict students' test performance, the *p*-value threshold was selected as 0.15 rather than 0.05 as the feature selection stop criteria (Faraway, 2002). Therefore, the number of pauses, as the third variable in Robust LMM, was also selected for the model during backward selection (see Table 11). In addition, using it in the model decreased RMSE. However, using other variables caused the *p*-value to exceed the 0.15 threshold, hence the model

Table 10 Comparison of three features and eight features with Robust LMM

No	Model	*N* (Observations)	RMSE	MAE	SD	Z	*p*
1	RLMM with 3 features	42	15.11	11.27	10.19	462	0.185
2	RLMM with 8 features	42	16.40	12.66	10.54		

Table 11 Fixed effects of final Robust LMM

Rank	Variable	Estimate	SE	*t*	*p*
0	Intercept	79.83	3.78	21.14	0.00
1	Mean Backward Speed	-162.14	51.48	-3.15	0.00
2	Percentage of Backwards	32.53	11.95	2.72	0.01
3	Number of Pauses	0.26	0.13	1.96	0.06

* $VIF_1 = 1.09$, $VIF_2 = 1.10$, $VIF_3 = 1.07$

presented in Table 11 was selected for the Robust LMM. Table 11 presents the fixed effects summary for the final Robust LMM model.

4.2.2 Comparing XGBoost and robust LMM with baseline (Nested stratified k-fold cross-validation)

Nested k-fold cross-validation was used where the original dataset was split into training and testing datasets, and stratified random sampling was used with five folds, five times, the same as in Step 1 of Section 4.2.1. In each iteration, one fold was set aside as an unseen testing dataset and the remaining four folds were used to validate the model, find the best model parameters, and for training purposes. The results for XGBoost showed that RMSE was 14.76 (M_{cv}=14.68, SD_{cv}=1.90). For Robust LMM, there was no need to perform hypermeter tuning. The results for Robust LMM showed that RMSE was 15.11 (M_{cv}=14.286, SD_{cv}=3.73).

However, the RMSE scores for XGBoost and Robust LMM are not meaningful without the use of a baseline for comparative purposes. Therefore, their results were each compared with a baseline model that predicted the test performance as the mean test performance scores calculated from the training dataset for all the data points in the test dataset.

A Friedman test (aka Friedman's ANOVA), a statistical analysis method used for comparing means of several related groups having non-parametric distributions (Field, 2018), was conducted to compare the XGBoost and Robust LMM performances with that of the baseline. The results revealed a significant difference ($\chi^2(2)=12.905$, $p=0.002$) (see Table 12). Wilcoxon-Bonferroni post hoc tests were conducted for pairwise comparison. The comparisons showed that both XGBoost (RMSE=14.76, MAE=11.50, SD=9.36) and Robust LMM (RMSE=15.11, MAE=11.27, SD=10.19) performances were significantly higher than the baseline model (RMSE=18.12, MAE=15.51, SD=9.48), p(XGBoost—Baseline)=0.024 and p(Robust LMM—Baseline)=0.009. However, there was no significant difference established between XGBoost and Robust LMM, p(XGBoost—Robust LMM)=1.000.

RMSE percentage (aka relative RMSE or normalized RMSE) was also used for comparison (see Tables 12, 13, and 14) (Jamieson et al., 1991; Jeong et al., 2014; Li et al., 2013):

$$\%RMSE = \frac{RMSE}{\overline{X}} \times 100$$

Table 12 Comparison of XGBoost, Robust LMM, and baseline (random split)

No	Model	N (Observations)	%RMSE	RMSE	MAE	SD	X^2	p
1	XGBoost	42	17.20	14.76	11.50	9.36	12.905	0.002
2	Robust LMM	42	17.61	15.11	11.27	10.19		
3	Baseline	42	21.12	18.12	15.51	9.48		

Table 13 Comparison of XGBoost, Robust LMM, and baseline model (one person leave out)

No	Model	*N* (Observations)	%RMSE	RMSE	MAE	SD	X^2	*p*
1	XGBoost	42	18.33	15.73	11.63	10.71	8.143	0.017
2	Robust LMM	42	17.79	15.26	11.52	10.14		
3	Baseline	42	22.26	19.10	16.60	9.56		

Table 14 Comparison of XGBoost, Robust LMM, and baseline (one video leave out)

No	Model	*N* (Observations)	%RMSE	RMSE	MAE	SD	X^2	*p*
1	XGBoost	42	19.86	17.04	11.99	12.25	11.423	0.003
2	Robust LMM	42	21.26	18.24	13.61	12.29		
3	Baseline	42	27.79	23.85	18.76	14.90		

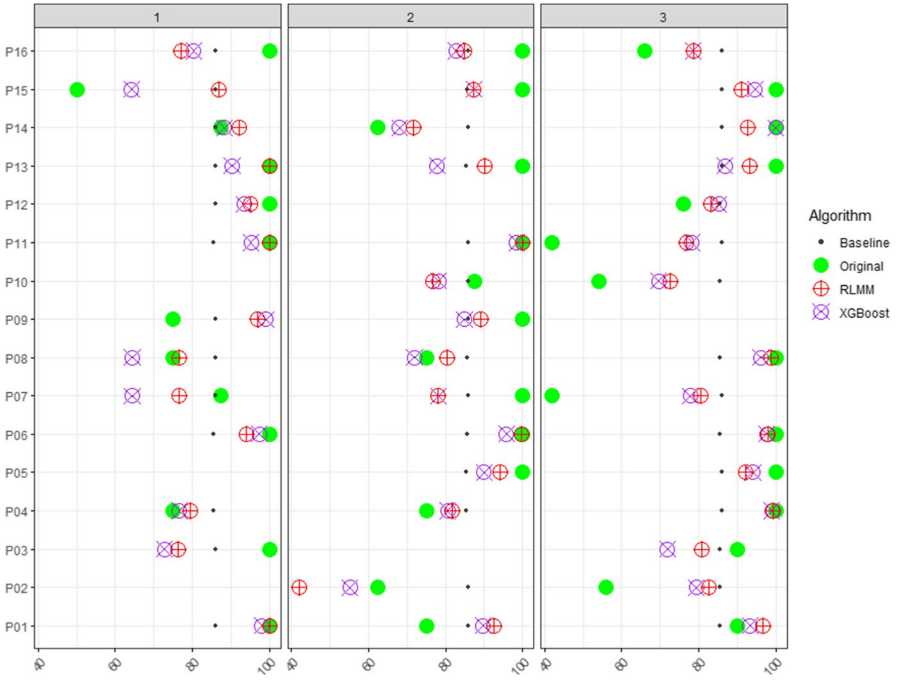

Fig. 2 Predictions of students' test performance (based on nested stratified 5-fold cross validation) for 42 observations

Figure 2 shows the test performance predictions for each student on each video. Both the XGBoost and Robust LMM models were considered good at predicting high values in a "nested 5-fold cross-validation" evaluation scheme. However, the models were shown to predict test performance of 70 or higher, except for P07, P08, P15 in Video 1 and P02 in Video 2, which had generally among the lowest scores in the related quizzes. Even where the models predicted test performance with low scores, they were not as low as the actual test performance scores. Similarly, P02, P07, P10, and also P11 received the lowest scores in the third video. Even if the error rates were considered high for each of them, the models' predictions were generally seen as lower for those participants with higher-end scores. This result may be explained by the limited prediction range of the models.

4.2.3 Comparison of XGBoost and robust LMM with baseline (One Person Leave Out)

This evaluation scheme aimed to measure the generalization capability of the proposed models for predicting new students' test performance using the existing video watching behaviors of other students. One student's data belonging to all three videos were left out for testing, with the remainder used for training and validation purposes. This process was conducted 16 times, which resulted in a total of 42 observations (videos). The baseline model was constructed to predict a new student's test performance, per quiz, based on the mean test performance of the other students (calculated using the training dataset). The results for XGBoost showed that RMSE was 15.73 ($M_{cv} = 13.60$, $SD_{cv} = 8.60$), whilst for Robust LMM the RMSE value was 15.26 ($M_{cv} = 13.37$, $SD_{cv} = 7.75$).

The Friedman test showed that a significant difference was found between the models ($\chi^2(2) = 8.143$, $p = 0.017$) (see Table 13). Wilcoxon-Bonferroni post hoc tests showed that both XGBoost (RMSE = 15.73, MAE = 11.63, $SD = 10.71$) and Robust LMM (RMSE = 15.26, MAE = 11.52, $SD = 10.14$) performances were significantly higher than the baseline according to the "mean" strategy (RMSE = 19.10, MAE = 16.60, $SD = 9.56$), p(XGBoost—Baseline) = 0.045 and p(Robust LMM—Baseline) = 0.000. However, there was no significant difference found between XGBoost and Robust LMM, p(XGBoost—Robust LMM) = 1.000.

As can be seen in Fig. 3, the predictions of each model were slightly worse than the "nested stratified 5-fold cross-validation" evaluation scheme. Both models produced high test performance scores with a high degree of accuracy; nevertheless, XGBoost was the more successful in predicting low scores, whilst less successful in predicting high scores than Robust LMM. Since the number of high scores was more than the number of low scores, Robust LMM appears to have provided better prediction scores. On the other hand, it was not easy to predict test performance for those students who had zero interaction or who had a low number of backward click interactions. For example, P02, P03, P07, and also P11 did not perform any backward click interactions when watching the third video. As a result, calculating their test performance was difficult since the number/percentage of backward clicks and their mean backward speed were all zero.

4.2.4 Comparing XGBoost and robust LMM with baseline (One Video Leave Out)

This evaluation scheme aimed to measure the generalization capability of the models for predicting student quiz grades after having watched a new video by utilizing their previous video watching behaviors. The students' video watching behaviors

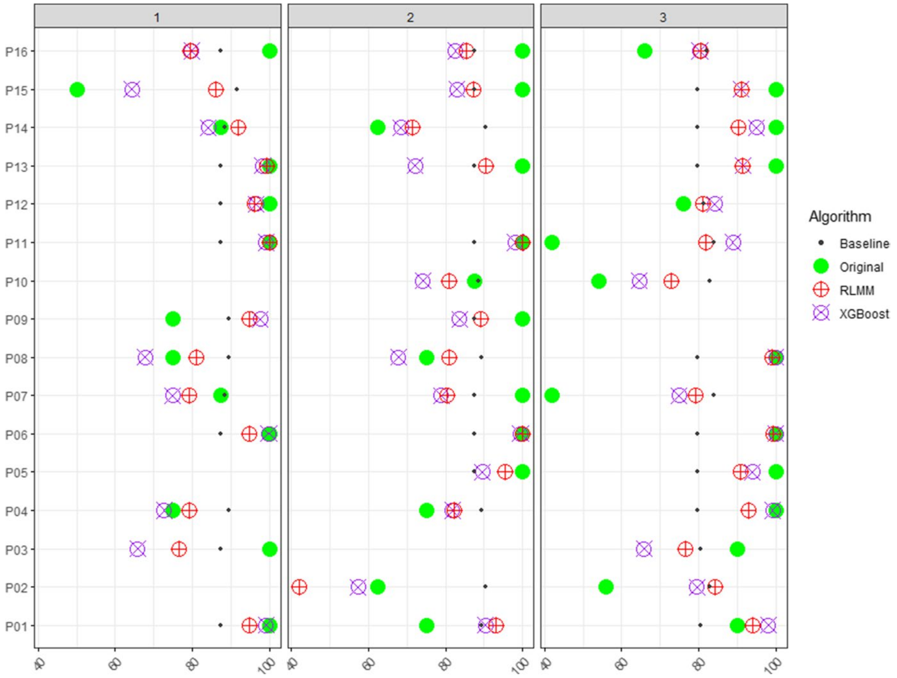

Fig. 3 Predictions of students' test performance (based on one person leave out) for 42 observations

for two previous videos were used for the training dataset, with one "new" video used for testing purposes. This process was performed three times, since there were three video lectures and therefore resulted in a total of 42 observations. The baseline models were constructed using the mean of the previous quiz grades for each student. Cross-validation was performed for XGBoost using the parameters and values presented in Table 5. The results for XGBoost showed that RMSE was 17.04 (M_{cv} = 16.57, SD_{cv} = 3.92), whilst for Robust LMM the RMSE value was 18.24 (M_{cv} = 17.82, SD_{cv} = 4.09). The comparative results presented in Table 14 show that the XGBoost model outperformed both the Robust LMM and the baseline models.

The Friedman test showed that a significant difference was found between the models based on the "one video leave out" test scenario ($\chi^2(2) = 11.423$, $p = 0.003$) (see Table 14). Wilcoxon-Bonferroni post hoc tests showed that both the XGBoost (RMSE = 17.04, MAE = 11.99, *SD* = 12.25) and the Robust LMM (RMSE = 18.24, MAE = 13.61, *SD* = 12.29) performances were significantly higher than that of the baseline model (RMSE = 23.85, MAE = 18.76, *SD* = 14.90), *p*(XGBoost—Baseline) = 0.006 and *p*(Robust LMM—Baseline) = 0.024. However, there was no significant difference found between XGBoost and Robust LMM, *p*(XGBoost—Robust LMM) = 1.000.

Figure 4 shows the individual prediction of the models based on the "one video leave out" evaluation scheme. The success rate seen from this evaluation scheme was lower than for both the "random split" and "one person leave out" schemes. XGBoost and Robust LMM appear to be more successful when students receive

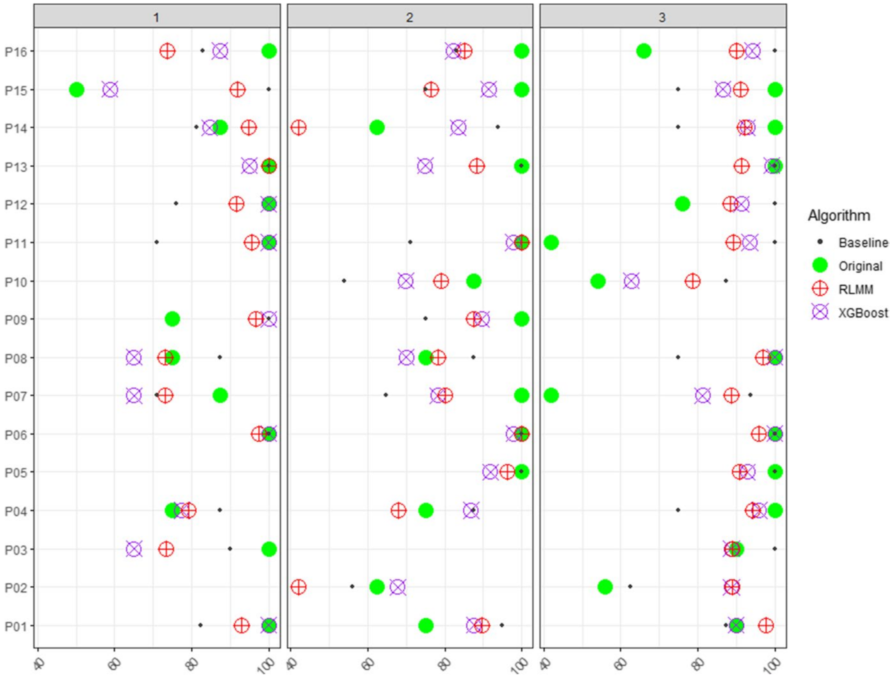

Fig. 4 Predictions of students' test performance (based on one video leave out) for 42 observations

high scores rather than low scores, as seen in the previously examined schemes. Correspondingly, it is not easy to predict the students' test performance based on the "one video leave out" evaluation scheme when students receive test performance scores far removed from the general tendency. For example, P07 and P11 both obtained a very low score in the quiz for the third video, even though both XGBoost and Robust LMM produced a lower error rate when compared to the baseline model. In addition, the number of student video lectures used within the construction of a model can play an important role in its prediction success. Therefore, the success of the models based on the "one video leave out" scenario would most probably increase if the number of videos increased.

The results based on the three evaluation validation strategies (Random Split, One Person Leave Out, and One Video Leave Out) showed that both XGBoost and Robust LMM performances were significantly higher than the baseline models.

5 Discussion and conclusion

In this study, the relationship between students' video clickstream behaviors and their test performance in both a midterm exam and multiple quizzes was investigated. The study's results showed that the total number of interactions had a significant and positive relationship with the students' subsequent test performance. That is, the more a student performed clickstream interactions when viewing video lectures, the higher their test performance was found to be. This is an important evidence that students need learner-content interaction in order to be successful during online courses (Zimmerman, 2012). The results also showed a negative correlation between the mean of backward speed and students' test performance. This indicates that clicking backward to the closest point demonstrates that the students had a desire to learn (i.e., to go back and review something they perhaps missed or did not fully understand), whereas clicking backward to a more remote point indicates that the students did not focus on their learning. The basic interaction types of pause click, backward click, and forward click were also investigated in detail, and the results showed that the numbers of pauses and backwards had a positive relation to the students' test performance. The findings obtained during the first experiment, which was based on a single measure, provided limited features; however, that analysis was then further supported by the second experiment which was based on repeated measures, and involved more detailed features and more in-depth analysis.

In summary, those students who performed high numbers of clickstream activities, slow backward speed, or those who preferred backward clicks or pauses received higher test performance scores than those students who performed a low number of clickstream activities, a fast backward speed, or those who preferred forward clicking when watching video lectures. These findings help to provide an informed perspective as to how best to design video-based lectures, and to help guide instructors as to which student behaviors are considered the most important in online learning. Behaviorism, one of the more popular learning theories, emphasizes that all behaviors can be gained through interactions with the environment (Graham, 2019). Accordingly, interactive videos can trigger desired behaviors by providing the appropriate level of interaction and the opportunity to do so. Especially, the use of embedded questioning techniques within interactive videos can help to improve students' learning and the efficiency of the time spent viewing video lectures (Vural, 2013). They can also contribute to a mental review or replay of video lectures as a means to achieving specified learning goals (Van der Meij & Böckmann, 2021). Therefore, to increase the number of backward clicks, pauses, high volume clicking, and slow backward speed, we suggest that *interactive questions (to promote high volume clicking) that provide students with the opportunity and need to search for and process information (for high numbers of backward clicks or pauses) can be added/embedded periodically (for slow backward speed) to the design of video-based lectures*. This approach could also help students learn by encouraging them to practice what they learn through interactive and inquisitive questioning.

Finally, the features that may be best used to predict students' test performance were investigated during the study. The XGBoost algorithm, one of the most powerful algorithms used for small-sized study samples, and the Robust Linear Mixed Model (RLMM), one of the most suitable algorithms for repeated measures, were used to compare the results, both with baseline models and with each other. While the XGBoost algorithm was shown to better predict students' test performance based on the "nested stratified 5-fold cross-validation" and "one video leave out" approaches, the Robust LMM gave better results when the model was validated based on the "one person leave out" approach. However, no significant statistical difference was found to exist between them. Therefore, both algorithms (XGBoost and Robust LMM) may be used for repeated measures collected from a small-sized sample. In addition, they both performed better than the baseline models. Therefore, we can conclude that students' test performance can be predicted with an error in the range of approximately 15 to 20 points (%RMSE) over a 100-point range. The obtained RMSE percentage reveals that students' test performance can be successfully predicted using video clickstream interactions modeled either with the XGBoost or Robust LMM since a value of 15–20 is considered good for %RMSE in prediction (Li et al., 2013).

Mean of backward speed, number of pauses, and number/percentage of backward clicks are seen as the most important indicators that can be used in predicting success or failure in the models. These indicators are compatible with the findings reported in the literature, even though none have previously been explicitly mentioned in any single study in terms of predicting students' learning or test performance. Lan et al. (2017) showed that the number of pauses and rewinds (i.e., backward clicks) were among the features that promoted learner engagement the most, which can be said to implicitly show consistency with the findings of the current study, as engagement positively contributes to the students' learning performance. Furthermore, Hasan et al. (2020) indicated that pausing videos many times or progressing slowly through videos increases the probability of students successfully passing a course. Lu et al. (2018) found that the number of backward clicks per week was also a relevant factor in predicting learning performance. This also justifies our findings from the current study, even though the number of backward clicks was not used for each video as a feature in Lu et al.'s (2018) study.

All of these findings suggest candidate potential algorithms for modeling and the use of video interaction features in different ways, but that the use of single scores for multiple video lectures together with different algorithms does not necessarily provide the sole means to predict grade outcomes of learners (Solli et al., 2018). However, these algorithms and features may be used for the implementation of timely interventions to increase the students' successes and to decrease their rate of failure (Van Goidsenhoven et al., 2020).

The existing studies in the literature that have been performed in the sphere of higher education have either not directly indicated the significance of certain video-related features (Hasan et al., 2020), or have remained only at the descriptive level (Yoon et al., 2021). Viberg et al. (2018) also emphasized that most studies focusing on learning analytics in the higher education context adopt a descriptive approach; a situation seen more clearly in the study of video analytics, which is a sub-field of learning analytics. Moreover, little is known about how to make best use of predictive learning analytics in the higher education context (Herodotou et al., 2019). Therefore, the current study contributes to the literature in terms of the usage of predictive analytics in video-based learning since it used a predictive approach, and which was shown to assist in improving educational quality by revealing critical situations (Doleck et al., 2020). With the help of predictive analytics, both educational institutions and course instructors can make better, more informed decisions and act accordingly based on actual data (Daniel, 2015).

In addition, the current study solely used the most basic of video clickstream behaviors, which require no additional information to be collected, and explained step-by-step how to analyze these behaviors. The study's results demonstrated that with just a few significant variables, the test performance of students can be predicted with a high degree of accuracy. For the first time in the related literature, the current study employed experiments that were conducted comprehensively with three different evaluation schemes. The results of these experiments revealed that the variables –number of pauses, mean backward speed, number of backwards, and percentage of backwards– were effective in predicting the performance of both new student users on existing videos or existing users on new videos.

The findings of the current study may therefore be considered as beneficial for all stakeholders working in the field of education. Firstly, students are the most important stakeholder group since the overriding priority in education is to deliver effective student learning. The study revealed which behaviors are the best indicators of measuring success in video-based learning. These indicators can be used for the early detection of problems regarding video-based materials or students' learning (i.e., the potential for failure). Then, any necessary revisions to course materials or interventions may be developed and implemented in a timely manner. In addition, instructors and MOOC designers can consider revising their lectures to increase the positive video clickstream behaviors of their learners. Finally, researchers or video-based learning platform developers can utilize the proposed educational data mining approach in order to create smart learning systems.

The results of the current study may also be seen as helpful for MOOC platforms that offer several courses, as well as for few courses taken by small groups, which make use of video lectures as supporting materials or as the sole teaching medium. In particular, the approach demonstrated in this study may be applied to any higher education course that adopts video-based learning, since it does not necessarily rely upon a large-sized sample to be effective.

In conclusion, the findings of the current study may be summarized as follows:

- Video clickstream behaviors can be used to detect students who are at risk of failure.
- The approach introduced in this study may also be applied to small classes with only limited numbers of students.
- The study has shown which video clickstream behaviors are considered important in the prediction of students' academic performance.
- The findings of the study can help teachers to better design their lectures.
- Applying the findings of the study can enable students to learn more effectively.
- The findings of the study can assist both researchers and developers in the application and evaluation of suitable educational data mining approaches that utilize video clickstream interactions.

6 Limitations and future work

Even though this study presents important findings on how to utilize video clickstream data to predict students' test performance and makes suggestions for designing instructional videos, it has a notable limitation: the study was conducted with a small number of participants. One of the most important reasons for this was there being only one suitable course whose instructor was also willing to work within the prescribed experimental study settings over two different semesters. In order to deal with this limitation, educational data mining techniques suited to a small sample were employed. Importantly, the total number of interactions performed by the study's participants were considered adequate in number so as to be able to legitimately investigate the behaviors.

In addition, each of the videos selected for analysis was approximately 30 min in duration, and were otherwise identical, albeit with different verbalized topics. The results of the study may therefore be considered generalizable to other video lectures of similar length, and that also present verbalized topics. However, in order to generalize the results to short videos or quantitative videos, additional research would be required.

In a future work, we are planning to investigate the reasons behind certain clicking behaviors. Having greater knowledge about students' video clickstream behaviors, and the reasons behind them, could provide valuable insight into learners' academic performance such as what click rates are the most effective. If the reasons are known by researchers and instructors, the possibility to predict students' test performance will likely increase. Therefore, we aim to investigate video clickstream behaviors and clickstream reasoning together in order to better predict students' test performance.

Moreover, we also see the benefits of using advanced data mining on a specific type of instructional material (video) upon which to base learning analytics. Analytical approaches can also be used and investigated specifically by future researchers for each type of learning object: e-books, audio, presentations, animations, and simulations, which may lead to the rise of a different sub-field of learning analytics such as reading analytics or e-book-based learning analytics.

Acknowledgements This research was carried out in the scope of the first author's doctoral dissertation under the supervision of the second and third authors.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Research involving human participants This research was granted with an approval from Middle East Technical University's Ethics Committee (Protocol No: 189-ODTÜ-2019).

Conflict of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Akçapınar, G., & Bayazıt, A. (2018). Investigating video viewing behaviors of students with different learning approaches. *Turkish Online Journal of Distance Education, 19*(4), 116–125. https://doi.org/10.17718/tojde.471907

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance inflation factor: As a condition for the inclusion of suppressor variable(s) in regression analysis. *Open Journal of Statistics, 5*(7), 754–767. https://doi.org/10.4236/ojs.2015.57075

Atapattu, T., & Falkner, K. (2018). Impact of lecturer's discourse for student video interactions: Video learning analytics case study of MOOCs. *Journal of Learning Analytics, 5*(3), 182–197. https://doi.org/10.18608/jla.2018.53.12

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology, 8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Brinton, C. G., Buccapatnam, S., Chiang, M., & Poor, H. V. (2015). *Mining MOOC Clickstreams: On the Relationship Between Learner Video-Watching Behavior and Performance*. https://arxiv.org/abs/1503.06489

Brooks, C., Thompson, C., & Greer, J. (2013). Visualizing lecture capture usage: A learning analytics case study. In *Proceedings of the Workshop on Analytics on Video-based Learning (WAVe 2013)*, *Vol. 983,* (pp. 9–14). CEUR. http://ceur-ws.org/Vol-983/paper3.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

Chen, T., He, T., & Benesty, M. (2015). *xgboost: eXtreme Gradient Boosting*. R Package Version 0.4–0, 0–4. https://cran.microsoft.com/snapshot/2015-10-20/web/packages/xgboost/xgboost.pdf

Chorianopoulos, K., & Giannakos, M. (2013). Merging learner performance with browsing behavior in video lectures. In *Proceedings of the Workshop on Analytics on Video-based Learning (WAVe 2013)*, *Vol. 983,* (pp. 38–42). CEUR. http://ceur-ws.org/Vol-983/paper9.pdf

Chorianopoulos, K., Leftheriotis, I., & Gkonela, C. (2011). SocialSkip: Pragmatic understanding within web video. In *EuroITV'11 – Proceedings of the 9th European Interactive TV Conference* (pp. 25–28). ACM. https://doi.org/10.1145/2000119.2000124

Daniel, B. (2015). Big Data and analytics in higher education: Opportunities and challenges. *British Journal of Educational Technology, 46*(5), 904–920. https://doi.org/10.1111/bjet.12230

De Boer, J., Kommers, P. A. M., & De Brock, B. (2011). Using learning styles and viewing styles in streaming video. *Computers and Education, 56*(3), 727–735. https://doi.org/10.1016/j.compedu.2010.10.015

Demidenko, E. (2013). *Mixed models: Theory and applications with R*. Wiley. https://doi.org/10.1002/9781118651537

Dimitrova, V., Mitrovic, A., Piotrkowicz, A., Lau, L., & Weerasinghe, A. (2017). Using learning analytics to devise interactive personalised nudges for active video watching. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization – UMAP 17* (pp. 22–31). ACM. https://doi.org/10.1145/3079628.3079683

Dissanayake, D., Perera, T., Elladeniya, C., Dissanayake, K., Herath, S., & Perera, I. (2018). Identifying the learning style of students in MOOCs using video interactions. *International Journal of Information and Education Technology, 8*(3), 171–177. https://doi.org/10.18178/ijiet.2018.8.3.1029

Doleck, T., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: A comparison of deep learning frameworks. *Education and Information Technologies, 25*(3), 1951–1963. https://doi.org/10.1007/s10639-019-10068-4

El Aouifi, H., El Hajji, M., Es-Saady, Y., & Douzi, H. (2021). Predicting learner's performance through video sequences viewing behavior analysis using educational data-mining. *Education and Information Technologies, 26*, 5799–5814. https://doi.org/10.1007/s10639-021-10512-4

Faraway, J. J. (2002). Variable selection. In *Practical Regression and Anova using R* (pp. 124–133). Cran Microsoft. https://cran.microsoft.com/snapshot/2015-03-01/doc/contrib/Faraway-PRA.pdf

Field, A. P. (2018). *Discovering statistics using IBM SPSS statistics*. Sage.

Fox, J. (2002). *Linear mixed models*. Sage.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Gardner, J., & Brooks, C. (2018). Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction, 28*(2), 127–203. https://doi.org/10.1007/s11257-018-9203-z

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2015). Making sense of video analytics: lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course. *International Review of Research in Open and Distributed Learning, 16*(1), 260–283. https://doi.org/10.19173/irrodl.v16i1.1976

Giannakos, M. N., Chorianopoulos, K., Ronchetti, M., Szegedi, P., & Teasley, S. D. (2013a). Expanding horizons and envisioning the future of analytics on video-based learning. In *Proceedings of the Workshop on Analytics on Video-based Learning (WAVe 2013), Vol. 983,* (pp. 1–6). CEUR. http://ceur-ws.org/Vol-983/paper1.pdf

Giannakos, M. N., Chorianopoulos, K., Ronchetti, M., Szegedi, P., & Teasley, S. D. (2013b). *Workshop on analytics on video-based learning (WAVe 2013).* http://ceur-ws.org/Vol-983/WAVe2013-Proceedings.pdf

Giannakos, M. N., Chorianopoulos, K., & Chrisochoides, N. (2014). Collecting and making sense of video learning analytics. In *2014 IEEE Frontiers in Education Conference (FIE)* (pp. 1–7). IEEE. https://doi.org/10.1109/FIE.2014.7044485

Giannakos, M. N., Sampson, D. G., Kidziński, L., & Pardo, A. (2016). Enhancing video-based learning experience through smart environments and analytics. In *Proceedings of the LAK 2016 Workshop on Smart Environments and Analytics in Video-Based Learning*, *Vol. 1579* (pp. 1–6). CEUR. http://ceur-ws.org/Vol-1579/paper5.pdf

Graham, G. (2019). Behaviorism. In E. N. Zalta (Ed.), *The {Stanford} Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/behaviorism/

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2015). Predicting students performance in educational data mining. In *Proceedings of International Symposium on Educational Technology (ISET)* (pp. 125–128). IEEE. https://doi.org/10.1109/ISET.2015.33

Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences, 10*(11), Article 3894. https://doi.org/10.3390/app10113894

Herodotou, C., Rienties, B., Verdin, B., & Boroowa, A. (2019). Predictive learning analytics 'at scale': Guidelines to successful implementation in Higher Education based on the case of the Open University UK. *Journal of Learning Analytics, 6*(1), 85–95. https://doi.org/10.18608/jla.2019.61.5

Hussain, A., Khan, M., & Ullah, K. (2022). Student's performance prediction model and affecting factors using classification techniques. *Education and Information Technologies*, 1-18. https://doi.org/10.1007/s10639-022-10988-8

Ilioudi, C., Giannakos, M. N., & Chorianopoulos, K. (2013). Investigating differences among the commonly used video lecture styles. In *Proceedings of the Workshop on Analytics on Video-based Learning (WAVe 2013), Vol. 983* (pp. 21–26). CEUR. http://ceur-ws.org/Vol-983/paper5.pdf

Jamieson, P. D., Porter, J. R., & Wilson, D. R. (1991). A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. *Field Crops Research, 27*(4), 337–350. https://doi.org/10.1016/0378-4290(91)90040-3

Jeong, H., Jang, T., Seong, C., & Park, S. (2014). Assessing nitrogen fertilizer rates and split applications using the DSSAT model for rice irrigated with urban wastewater. *Agricultural Water Management, 141*, 1–9. https://doi.org/10.1016/j.agwat.2014.04.009

Kleftodimos, A., & Evangelidis, G. (2014). Exploring student viewing behaviors in online educational videos. In *Proceedings - IEEE 14th International Conference on Advanced Learning Technologies* (ICALT 2014) (pp. 367–369). IEEE. https://doi.org/10.1109/ICALT.2014.109

Kleftodimos, A. (2016). *Video Based Learning Analytics: using open source tools and open Internet resources for building interactive video based learning environments that support learning analytics* [Doctoral dissertation, University of Macedonia, Thessaloniki, Greece]. https://thesis.ekt.gr/thesisBookReader/id/43519

Kloft, M., Stiehler, F., Zheng, Z., & Pinkwart, N. (2014). Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs* (pp. 60–65). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-4111

Koller, M. (2016). robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software, 75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Lan, A. S., Brinton, C. G., Yang, T.-Y., & Chiang, M. (2017). Behavior-based latent variable model for learner engagement. In *Proceedings of the 10th International Conference on Educational Data Mining (EDM 2017)* (pp. 64–71). EDM. http://educationaldatamining.org/EDM2017/proc_files/proceedings.pdf

Lau, K. H. V., Farooque, P., Leydon, G., Schwartz, M. L., Mark, R., & Moeller, J. J. (2018). Using learning analytics to evaluate a video-based lecture series. *Medical Teacher, 40*(1), 91–98. https://doi.org/10.1080/0142159X.2017.1395001

Li, M. F., Tang, X. P., Wu, W., & Liu, H. B. (2013). General models for estimating daily global solar radiation for different solar radiation zones in mainland China. *Energy Conversion and Management, 70*, 139–148. https://doi.org/10.1016/j.enconman.2013.03.004

Li, N., Kidziński, Ł, Jermann, P., & Dillenbourg, P. (2015). MOOC video interaction patterns: What do they tell us? In G. Conole, T. Klobučar, C. Rensing, J. Konert, & E. Lavoué (Eds.), *Design for Teaching and Learning in a Networked World. EC-TEL 2015. Lecture Notes in Computer Science* (pp. 197–210). Springer. https://doi.org/10.1007/978-3-319-24258-3_15

Li, X., Xie, L., & Wang, H. (2016). Grade prediction in MOOCs. In *2016 IEEE Intl Conference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)* (pp. 386–392). IEEE. https://doi.org/10.1109/CSE-EUC-DCABES.2016.213

Lu, O., Huang, A., Huang, J., Lin, A., Ogata, H., & Yang, S. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Journal of Educational Technology & Society, 21*(2), 220–232. https://doi.org/10.2307/26388400

Mbouzao, B., Desmarais, M. C., & Shrier, I. (2020). Early prediction of success in MOOC from video interaction features. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education. AIED 2020. Lecture Notes in Computer Science* (Vol. 12164, pp. 191–196). Springer. https://doi.org/10.1007/978-3-030-52240-7_35

Mirriahi, N., & Vigentini, L. (2017). Analytics of learner video use. In C. Lang, G. Siemens, A. Wise, & D. Gašević (Eds.), *Handbook of learning analytics* (pp. 251–267). Solar. https://doi.org/10.18608/hla17.022

Mubarak, A. A., Cao, H., & Ahmed, S. A. M. (2021). Predictive learning analytics using deep learning model in MOOCs' courses videos. *Education and Information Technologies, 26*(1), 371–392. https://doi.org/10.1007/s10639-020-10273-6

Ronchetti, M. (2013). Videolectures ingredients that can make analytics effective. In *Proceedings of the Workshop on Analytics on Video-based Learning (WAVe 2013), Vol. 983* (pp. 15–20). CEUR-WS. http://ceur-ws.org/Vol-983/paper4.pdf

Rose, C., & Siemens, G. (2014). *EMNLP 2014 workshop on analysis of large scale social interaction in MOOCs*. https://aclanthology.org/W14-4100.pdf

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution, 11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Seidel, N. (2017). Analytics on video-based learning. A literature review. In C. Ullrich & M. Wessner (Eds.), *Proceedings of DeLFI and GMW Workshops 2017*. CEUR. http://ceur-ws.org/Vol-2092/paper4.pdf

Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science, 72*, 414–422. https://doi.org/10.1016/j.procs.2015.12.157

Shi, X., Wong, Y. D., Li, M. Z. F., Palanisamy, C., & Chai, C. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. *Accident Analysis and Prevention, 129*, 170–179. https://doi.org/10.1016/j.aap.2019.05.005

Sinha, T., Jermann, P., Li, N., & Dillenbourg, P. (2014a). Your click decides your fate: Inferring information processing and attrition behavior from mooc video clickstream interactions. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1407.7131

Sinha, T., Li, N., Jermann, P., & Dillenbourg, P. (2014b). Capturing attrition intensifying structural traits from didactic interaction sequences of MOOC learners. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1409.5887

Solli, R., Aiken, J. M., Henderson, R., & Caballero, M. D. (2018). Examining the relationship between student performance and video interactions. In *Proceedings of the 2018 Physics Education Research Conference*. AAPT. https://doi.org/10.1119/perc.2018.pr.solli

Soni, A., Kumar, V., Kaur, R., & Hemavath, D. (2018). Predicting student performance using data mining techniques. *International Journal of Pure and applied Mathematics, 119*(SI 12), 221–227. https://acadpubl.eu/hub/2018-119-12/articles/7/1591.pdf.

Suthers, D. D., Verbert, K., Duval, E., & Ochoa, X. (2013). *LAK 2013: Third international conference on learning analytics and knowledge*. Leuven, Belgium, April 08-12, 2013. Association for Computing Machinery, Inc.

Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2018). Personality predictions based on user behavior on the Facebook social media platform. *IEEE Access, 6*, 61959–61969. https://doi.org/10.1109/ACCESS.2018.2876502

Ullrich, C., Shen, R., & Xie, W. (2013). Analyzing student viewing patterns in lecture videos. In *2013 IEEE 13th International Conference on Advanced Learning Technologies* (pp. 115–117). IEEE. https://doi.org/10.1109/ICALT.2013.38

Van der Meij, H., & Böckmann, L. (2021). Effects of embedded questions in recorded lectures. *Journal of Computing in Higher Education, 33*, 235–254. https://doi.org/10.1007/s12528-020-09263-x

Van Goidsenhoven, S., Bogdanova, D., Deeva, G., Vanden Broucke, S., De Weerdt, J., & Snoeck, M. (2020). Predicting student success in a blended learning environment. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge* (LAK '20) (pp. 17–25). ACM. https://doi.org/10.1145/3375462.3375494

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior, 89*, 98–110. https://doi.org/10.1016/j.chb.2018.07.027

Vural, Ö. F. (2013). The impact of a question-embedded video-based learning tool on e-learning. *Educational Sciences: Theory and Practice, 13*(2), 1315–1323.

Yang, T.-Y., Brinton, C. G., Joe-Wong, C., & Chiang, M. (2017). Behavior-Based Grade Prediction for MOOCs Via Time Series Neural Networks. *IEEE Journal on Selected Topics in Signal Processing, 11*(5), 716–728. https://doi.org/10.1109/JSTSP.2017.2700227

Yoon, M., Lee, J., & Jo, I. H. (2021). Video learning analytics: Investigating behavioral patterns and learner clusters in video-based online learning. *The Internet and Higher Education, 50*, 100806. https://doi.org/10.1016/j.iheduc.2021.100806

Yousef, A. M. F., Chatti, M. A., & Schroeder, U. (2014). Video-based learning: A critical analysis of the research published in 2003–2013 and future visions. In S. White (Ed.), *ELmL 2014: The Sixth International Conference on Mobile, Hybrid and On-Line Learning* (pp. 112–119). International Academy, Research, and Industry Association (IARIA).

Yu, C.-H., Wu, J., & Liu, A.-C. (2019). Predicting learning outcomes with MOOC clickstreams. *Education Sciences, 9*(2), Article 104. https://doi.org/10.3390/educsci9020104

Yürüm, O. R., Yıldırım, S., & Taşkaya-Temizel, T. (2022). An intervention framework for developing interactive video lectures based on video clickstream behavior: a quasi-experimental evaluation. *Interactive Learning Environments*, 1-16. https://doi.org/10.1080/10494820.2022.2042312

Zimmerman, T. D. (2012). Exploring learner to content interaction as a success factor in online courses. *International Review of Research in Open and Distance Learning, 13*(4), 152–165. https://doi.org/10.19173/irrodl.v13i4.1302

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Ozan Raşit Yürüm[1] · Tuğba Taşkaya-Temizel[2] · Soner Yıldırım[3]

[1] Distance Education Application and Research Center, İzmir Institute of Technology, İzmir, Turkey

[2] Department of Data Informatics, Middle East Technical University, Ankara, Turkey

[3] Department of Computer Education and Instructional Technology, Middle East Technical University, Ankara, Turkey