Video Features for Predicting Knowledge Gain in Search as Learning

Wolfgang Bitter[1], Anett Hoppe[1,2] and Ralph Ewerth[1,2]

[1]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]University of Marburg and hessian.AI – Hessian Center for Artificial Intelligence, Marburg, Germany

Abstract

While video platforms increasingly serve as primary learning resources during exploratory web searches, current approaches to predicting knowledge gain largely ignore video-specific features. This paper bridges this gap by examining how video interaction features (e.g., pausing, rewinding, forward navigation, viewing coverage) and video resource features (e.g., words per minute in speech transcripts, complex word ratios, and video file size density) correlate with learning outcomes. Using a publicly available dataset of 94 participants who engaged with educational videos during their search sessions, our analysis reveals that video interaction features, particularly those related to interaction frequency, are the strongest predictors of learning outcomes. Moreover, we analyze the influence of individual features on classification performance, revealing distinct relationships between different types of video interactions and knowledge gain. While our study is exploratory and based on a limited dataset, it provides valuable first insights and a foundation for future research on video-based learning behavior in search as learning settings. These insights can inform the design of adaptive learning systems that recognize and promote productive video engagement behaviors. To support future research, we release our feature extraction pipeline and analysis code[1].

Keywords

Search as Learning, Knowledge Gain Prediction, Video Learning, Video Interactions

1. Introduction

The internet has transformed how people acquire knowledge, with web searches playing a pivotal role in informal learning.

Educational videos have become increasingly important in addition to traditional textual resources. They offer learners an engaging way to process complex topics through rich visual and auditory elements. Platforms like YouTube are crucial tools in these learning journeys. They enable users to control the pace of their exploration and revisit challenging content sections.

The research area Search as Learning (SaL) investigates web search sessions with a learning intent [1]. Recently, research on SaL has made significant strides in understanding how web searches facilitate knowledge acquisition. A considerable body of work has explored the relationship between learning outcomes and both user behavior (e.g., query patterns, clickstreams) [2, 3, 4] and the properties of consumed resources (e.g., textual complexity, readability) [5, 6, 7, 8]. However, while these studies offer valuable insights into textual resources, videos—which are inherently multimodal and interactive—remain understudied in the context of SaL.

Educational videos are uniquely suited for learning because they combine multiple forms of information, as emphasized in the Cognitive Theory of Multimedia Learning (CTML)[9]. Interactions such as pausing, rewinding, and forward jumping allow learners to adapt content delivery to their needs, potentially enhancing comprehension and retention. Prior studies have examined how these interactions

[1]https://github.com/TIBHannover/sal_video_interactions

Disinformation, Misinformation and Learning in the Age of Generative AI: Joint Proceedings of the 1st International Workshop on Disinformation and Misinformation in the Age of Generative AI (DISMISS-FAKE'25) and the 4th International Workshop on Investigating Learning during Web Search (IWILDS'25) co-located with 18th International ACM WSDM Conference on Web Search and Data Mining (WSDM 2025)

wolfgang.bitter@tib.eu (W. Bitter); anett.hoppe@uni-marburg.de (A. Hoppe); ralph.ewerth@tib.eu (R. Ewerth)
0000-0003-1668-3304 (W. Bitter); 0000-0002-1452-9509 (A. Hoppe); 0000-0003-0918-6297 (R. Ewerth)

correlate with learning in controlled settings [10, 11], but there is limited research investigating their role in authentic web search contexts.

This paper addresses this gap by analyzing user interactions with videos during learning-oriented web searches. Specifically, we explore the extent to which interaction features can predict knowledge gain (KG), providing a deeper understanding of how video consumption contributes to learning in informal settings. The study focuses on two key research questions:

> *RQ1*: How do user interactions (e.g., pausing, rewinding) correlate with learning outcomes during video consumption in web search sessions?
>
> *RQ2*: To what extent can these interactions predict knowledge gain in such contexts?

To answer these questions, we use the publicly available *SaL-Lightning* dataset [12]. We developed a semi-automated approach for extracting detailed interaction logs from screen recordings. These logs capture user behaviors and are analyzed to uncover their relationships with knowledge outcomes. By focusing on videos as part of the broader SaL framework, our study expands the current understanding of multimodal learning resources and their role in web-based knowledge acquisition.

The remainder of the paper is structured as follows: Section 2 summarizes the recent research on SaL and video learning. Next, in Section 3, we explain how we extracted interaction logs from screen recordings and define the extracted video features. In Section 4, we describe the evaluation process and give insights into the relationship between video interaction and KG. Finally, in Section 5, we conclude the results and summarize the implications for future research.

2. Related Work

Search engines are increasingly used as learning tools. Therefore, it becomes imperative to design systems prioritizing knowledge acquisition [13]. The field of Search as Learning investigates user behavior and system design to enhance learning outcomes during web-based searches [14, 15].

Research in SaL has explored diverse aspects of learning-related interactions. For instance, Vakkari [16] identified features that reflect users' learning needs and their influence on knowledge acquisition during search activities. Similarly, Roy et al. [17] highlighted that learning is affected by prior knowledge, emphasizing the importance of modeling users' knowledge states and their evolution during the search process [18, 19, 20, 21].

Efforts to study factors influencing KG during the search can be broadly categorized into two research streams: a focus on (a) characteristics of web resources and (b) user behavior. For instance, Syed and Collins-Thompson [22] studied document-level features to improve learning outcomes, particularly for vocabulary acquisition. Ghafourian et al. [6] and Gritz et al. [5] explored readability metrics and textual complexity, demonstrating their impact on user behavior and KG prediction. Yu et al. [23, 7] utilized a wide range of features, including text and HTML statistics, to predict KG, while Otto et al. [24] investigated how multimedia features complement readability and linguistic factors in predicting learning outcomes. Recently, Gritz et al. [8] found a moderating influence of the visual complexity of web pages on learning outcomes.

Furthermore, research has shown that user behavior differs across more and less successful web searches. For example, input queries [2], navigation logs [3], and behavioral features such as time spent on pages or click patterns [4] have been linked to learning outcomes. These studies provide insights into how user interactions reflect and impact knowledge acquisition.

In the context of SaL, videos offer a multimodal learning experience that extends beyond traditional textual resources. Videos enhance comprehension and retention by combining visual and auditory elements, a principle supported by the Cognitive Theory of Multimedia Learning (CTML) [9]. User interactions, such as pausing, rewinding, and forwarding, are critical in learners' engagement with video content. For example, the segmenting effect described in CTML suggests that breaking multimedia content into smaller segments can reduce cognitive load and improve learning outcomes [25, 11]. Pausing

behavior, in particular, can reflect moments when learners process or integrate new information, aligning with points of high complexity or meaningful content structure [10, 26].

Despite the advancements in SaL and video learning research, the role of user interactions with videos within exploratory web searches remains relatively unexplored. Previous research has primarily focused on textual resources, leaving a gap in understanding how video-based interactions contribute to knowledge acquisition. Our work addresses this gap by examining video-specific user interaction features and their influence on KG prediction. This contributes to a more holistic understanding of learning in the context of SaL.

3. Methods

We developed systematic methods with three main components to investigate how video interactions and resource characteristics influence learning outcomes during web searches. First, we selected a dataset that captures both user interactions with educational videos and the measurement of knowledge gain (Section 3.1). Next, we implemented a semi-automatic approach to extract video interaction data from screen recordings (Section 3.2). Finally, we derived two sets of features: interaction features that quantify user engagement patterns and resource features that characterize video content properties (Section 3.3). This set of methods enables us to analyze how different aspects of video engagement correlate with knowledge acquisition during exploratory searches.

3.1. Rationale for Selection of the Dataset

After reviewing datasets for exploratory web searches from the literature [4, 27, 12], we decided to use the *SaL-Lightning* dataset [12]. This dataset proved optimal for our purposes for several reasons. First, it includes pre and post-test data necessary for measuring learning gains, unlike alternatives such as CoST [27]. Second, it captures diverse web navigation patterns with substantial video engagement, with $82\,\%$ of participants (94 of 114) accessing YouTube videos. This contrasts with other datasets such as Gadiraju et al. [4], where participants primarily accessed textual content like Wikipedia articles. While the published dataset includes standard user actions (clicks, scrolling), it lacks interactions with videos. We obtained access to the original screen recordings, enabling us to extract detailed video interaction data through semi-automated processing.

3.2. Extraction of Video Interactions from Screen Recordings

The manual screen recording annotation is time-consuming and error-prone, requiring continuous attention and precise temporal documentation. To address this, we developed a semi-automated approach, as depicted in Figure 1. The core idea involves fine-tuning and overfitting the object detection algorithm YOLO [28] and the OCR model TrOCR [29] on the study data to generate accurate interaction logs.

Using the provided timelines, we automatically extracted individual clips from the screen recordings, each representing a continuous sequence of a participant watching a YouTube video. Each clip was sampled at 10 frames per second. YOLO was used to detect the play/pause icon and the video playback position displayed on the interface. We fine-tuned YOLO using an initial training set of 25 manually annotated framed. Through iterative quality reviews, we addressed detected errors by expanding the training dataset, ultimately achieving reliable performance with 79 annotated frames.

The OCR algorithm TrOCR was applied to extract video timestamps. Similarly, misrecognized timestamps identified during quality checks were iteratively added to the training data, resulting in a total of 344 annotations. This process yielded interaction logs at a resolution of 10 timestamps per second, capturing whether the video was playing or paused.

The data were smoothed using a rolling maximum approach to address frame-to-frame inconsistencies in the detection results. We applied a seven-frame sliding window (three frames before and after each target frame) to determine the video timestamp and the playing status (paused/playing) through a

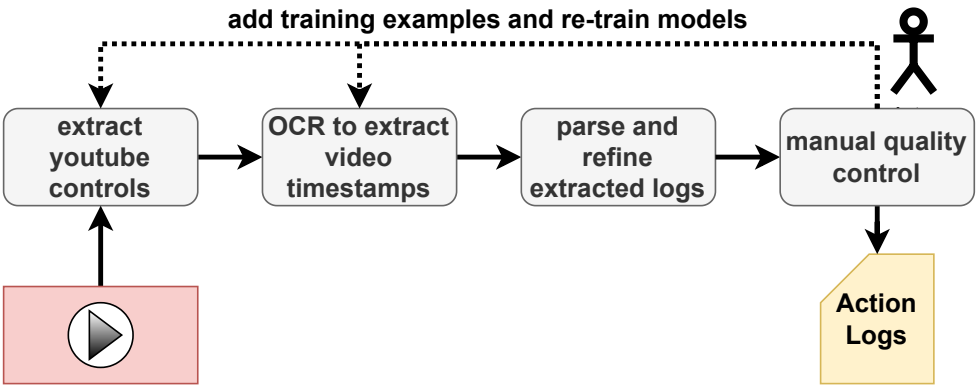

Figure 1: The figure depicts our general semi-automatic workflow to gather interaction log data: (1) YouTube controls in the screen recordings are extracted for the playing/paused symbol and the video timestamps using YOLO [28]; (2) the video timestamps get recognized by an OCR model [29]; (3) the video timestamps are parsed and the time series refined; (4) the results are manually and qualitatively judged and mistakes are added to the training data (whether for object detection or OCR).

majority vote. To validate the resulting interaction logs, we randomly sampled clips from different participants and manually cross-checked them against the original screen recordings until no further errors were identified.

3.3. Feature Calculation

The primary focus of this study is to investigate user interactions with videos during learning intended web search (see Section 3.3.1). However, we additionally experiment with video resource features, analyzing how the selected videos themselves can influence learning (see Section 3.3.2). Table 1 shows a complete list of extracted features.

3.3.1. Video Interaction Features

First, we define features based on accessing the videos (f_1-f_5) rather than actions on the video (e.g., pausing the video). Features f_1 and f_2 represent the number of videos and total dwell time on these videos. Next, f_3-f_5 indicate how much time the user spends on the video in relation to the video duration (e.g., $80\,\%$ of the videos on average). Features f_6 and f_7 reflect when the user interacts with the videos within the session or clip, which expresses whether a user primarily interacts with the videos relatively early or late. On the other hand, f_8 captures the timestamp of the earliest interaction across all watched videos. f_9-f_{12} count the absolute number of pauses, rewinds, seek forwards, and all video interactions per user across the whole search session:

$$f_9 = \sum_{v}^{V} pauses_v,\ f_{10} = \sum_{v}^{V} rewinds_v,\ f_{11} = \sum_{v}^{V} forwards_v,\ f_{12} = f_9 + f_{10} + f_{11}, \tag{1}$$

where V is the set of visited videos by a learner. Additionally, we capture the interaction rate by diving the total number of interactions by the total dwell time:

$$f_{13} = \frac{f_9}{DT},\ f_{14} = \frac{f_{10}}{DT},\ f_{15} = \frac{f_{11}}{DT},\ f_{16} = \frac{f_{12}}{DT},\ DT = \sum_{v}^{V} dwell\,time_v \tag{2}$$

where DT is the total dwell time per user on videos. These features give insights into how actively the users interact with the videos. $f_{17} - f_{19}$ measure the total duration of pauses, respectively, the total rewind and forward jump distance in seconds. f_{20} is the average pausing length, covering whether a

Table 1
Complete list of all calculated features.

	#	Description
video interaction features	f_1	Number of accessed YouTube videos
	f_2	Total sum of dwell times on YouTube videos
	f_3	Average percentage of the videos that have been viewed
	f_4	Highest percentage of the videos that have been viewed
	f_5	Total percentage of the videos that have been viewed
	f_6	Average time to interaction within a session
	f_7	Average time to interaction within a clip
	f_8	Earliest interaction within a clip
	f_9	Number of forward jump actions across a session
	f_{10}	Number of pauses across a session
	f_{11}	Number of rewind actions across a session
	f_{12}	Number of user interactions across a session
	f_{13}	Forward jumps per minute of total dwell time
	f_{14}	Pauses per minute of total dwell time
	f_{15}	Rewind actions per minute of total dwell time
	f_{16}	Interactions per minute of total dwell time
	f_{17}	Total sum of all pause duration
	f_{18}	Total sum of rewind distances
	f_{19}	Total sum of forward jump distances
	f_{20}	Ratio of total pausing duration and total dwell time
	f_{21}	Ratio of total rewind distance and total dwell time
	f_{22}	Ratio of total forward jump distance and total dwell time
	f_{23}	Average pausing duration
	f_{24}	Average rewind distance
	f_{25}	Average forward jump distance
	f_{26}	Ratio of the total rewinds to total forward jump actions
	f_{27}	Ratio of the total rewind distance to the total forward jump distance
	f_{28}	Ratio of the total forward jump actions to the sum of pauses and rewinds
	f_{29}	Relative time until interaction in relation to the total video duration
	f_{30}	Relative time until pause action in relation to the total video duration
	f_{31}	Relative time until rewind action in relation to the total video duration
	f_{32}	Relative time until forward jump action in relation to the total video duration
video resource features	f_{33}	Weighted average of characters in speech transcripts
	f_{34}	Weighted average of syllables in speech transcripts
	f_{35}	Weighted average of words in speech transcripts
	f_{36}	Weighted average of complex words in speech transcripts
	f_{37}	Weighted average of characters per minute in speech transcripts
	f_{38}	Weighted average of syllables per minute in speech transcripts
	f_{39}	Weighted average of words per minute in speech transcripts
	f_{40}	Weighted average of complex words per minute in speech transcripts
	f_{41}	Weighted average of video file size per frame per pixel

learner takes short or longer pauses. Further, f_{21} is the average rewind distance and reflects whether a user rewinds relatively small portions of the video or repeats whole videos. On the other hand, f_{22} is the average forward jump distance and indicates whether a user searches thoroughly or broadly for information. Similarly, $f_{23} - f_{25}$ measure the average pausing duration, rewind, and forward jump distance, independent of the dwell time. f_{26}-f_{28} measures the relationship between rewinds (and pauses) and forward jumps, indicating whether a user is profoundly engaging with videos or broadly scanning (e.g., searching for specific information). Finally, f_{29}-f_{32} reflect at which average percentage of the video the user performs an action (e.g., on average, after $20\,\%$ of the video a user pauses the video).

3.3.2. Video Resource Features

Using the tool `yt-dlp` [30], we downloaded each accessed video along with its associated metadata (e.g., video length, file size). Next, we used `Whisper` [31] (version `large-v3`) to get speech transcripts of the videos. Inspired by text complexity and readability research, we extracted the number of characters, syllables, words, and complex words from every speech transcript with the Python tool readability [32]. Moreover, similar to Gritz et al. [8], we used file size per frame per pixel as a proxy for visual complexity based on the principle that more complex visual content typically results in larger compressed file sizes.

Since every learner can access multiple videos, we calculate the average of the features per search session, weighted by the dwell time per video length:

$$F(x) = \frac{\sum_{v}^{V} x_v \cdot \frac{dwell\,time_v}{duration_v}}{\sum_{v=1}^{V} dwell\,time_v}, \tag{3}$$

where V is the set of videos accessed by a learner and $x \in$ {characters, syllables, words, complex words}. Subsequently, we define the features as follows:

$$f_{33} = F(\text{characters}), f_{34} = F(\text{syllables}), f_{35} = F(\text{words}), f_{36} = F(\text{complex words}). \tag{4}$$

We further divide all features by the total dwell time on videos, ensuring that our features capture interaction patterns rather than time spent on videos. Additionally, we then recalculate these features relative to video duration rather than total word count, yielding measures of speech rate (e.g., words per minute):

$$G(x) = \frac{\sum_{v=1}^{V} \frac{x_v}{duration_v} \cdot \frac{dwell\,time_v}{duration_v}}{\sum_{v}^{V} dwell\,time_v} \tag{5}$$

Again, we define the features as follows:

$$f_{37} = G(\text{characters}) f_{38} = G(\text{syllables}), f_{39} = G(\text{words}), f_{40} = G(\text{complex words}). \tag{6}$$

Finally, we use the file size of the `MP4 files`, normalized by the number of frames and per pixel:

$$f_{41} = \frac{\sum_{v}^{V} \frac{file\,size_v}{frames_v} \cdot \frac{dwell\,time_v}{duration_v}}{\sum_{v=1}^{V} dwell\,time_v}. \tag{7}$$

4. Evaluation

In this section, we assess the effectiveness of our features by analyzing their performance for the task of knowledge gain prediction. We begin by introducing the dataset used in our experiments (see Section 4.1), followed by a detailed explanation of the knowledge gain (see Section 4.2) and the evaluation metrics (see Section 4.3). For evaluation, we formulate the knowledge gain prediction as a classification task. In this regard, we present the experimental setup (see Section 4.4), including the classifiers, baselines, and feature selection methodology, and conclude with a discussion of the results (see Section 4.4.4) and an analysis of feature importance (see Section 4.5). This comprehensive evaluation aims to shed light on the role of video interactions in predicting learning outcomes and to identify the most relevant features contributing to classification performance.

4.1. Dataset

For our evaluation, we utilized the publicly available *SaL-Lightning* dataset, which focuses on exploratory web searches [12]. A total of 130 university students took part in the study, of which the data of 114 learners remained after filtering by the authors. Participants in this study were instructed to learn as much as they could about the generation of lightning and thunder within a time limit of 30 minutes. Still, they were allowed to finish whenever they wanted. Despite the time limit, the search was unrestricted,

meaning that any search engine and any web page on the Internet could be accessed, resulting in 808 unique visited URLs. To assess learning, the participants completed identical 10-question multiple-choice tests both one week before and immediately following the search sessions. Compensation was provided to all participants for their participation.

Since we are interested in the interactions with videos in this study, we filtered the participants according to the criterion that they accessed at least one YouTube video (N=94). The remaining participants were predominantly female (79 female, 15 male) and 22.8 ± 2.8 years old. On overage, the participants visited 4.5 ± 2.6 (1-17) videos for a total duration of $598\,\mathrm{s} \pm 338\,\mathrm{s}$ (82 s-1728 s). The participants had 14.5 ± 18.8 (0-135) interactions (pause, rewind, jump forward), while 7 did not interact with the videos at all.

4.2. Definition of Knowledge Gain

In recent works, knowledge gain was primarily measured as the difference between post-test and pre-test scores (correctly answered items). However, this does not consider the learner's confidence (e.g., guessing the answer). Therefore, we define a new measure to incorporate confidence. In the first step, we weigh the pre and post-tests with confidence.

$$pre,\, post = \sum_{i=1}^{n} (2 * correctness_i - 1) \cdot \frac{confidence_i}{3}, \tag{8}$$

where $correctness \in \{0, 1\}$ represents whether an answer was correct and $confidence \in \{0, 1, 2, 3\}$ the submitted confidence in the correctness for item i for n=10 items. This results in a score for each item between -1 (very confident but wrong) and 1 (very confident and correct). We combine pre and $post$ and define the knowledge gain as:

$$KG = max(\frac{1 + post - pre}{1 + n - pre}, 0). \tag{9}$$

We assume the learner's knowledge cannot decrease through a web search. Thus, the values can range between 0 and 1, where 0 means that nothing was learned and 1 that everything was correct in the post-test with full confidence.

Based on the literature, the actual values are less important than the classification of whether a web search leads to low, moderate, or high knowledge gain. Therefore, we define three categories—low, moderate, and high—and assign participants to these categories based on their KG as follows:

$$z(KG) = \frac{KG}{\sigma}, \tag{10}$$

where μ represents the mean and σ the standard deviation of the knowledge gains.

This results in the following distribution:

- **low:** 33 participants
- **moderate:** 25 participants
- **high:** 36 participants

4.3. Metrics

To evaluate the classification performance of the models and thus the predictive power of the features, we utilize the metrics Precision (p), Recall (r), F1-score (F_1), and Accuracy (acc)). These metrics are defined as follows:

$$p = \frac{TP}{TP + FP}, \quad r = \frac{TP}{TP + FN}, \tag{11}$$

$$F_1 = \frac{2 \cdot p \cdot r}{p + r}, \tag{12}$$

Table 2
Precision (p), recall (r), and F_1-scores (F_1) for the three classes *low*, *moderate*, and *high*. Additionally macro scores and micro accuracy (acc) for all classifiers and averaged. Results are grouped by baselines, video interaction features, video resource features, and the combination of both. Significant better results than baselines are highlighted in bold.

		low			moderate			high			macro			
	estimator	p	r	F_1	p	r	F_1	p	r	F_1	p	r	F_1	acc
baselines	majority	0.0	0.0	0.0	0.0	0.0	0.0	38.3	100.0	55.4	12.8	33.3	18.5	38.3
	stratified	25.3	23.0	24.1	26.0	31.2	28.4	35.9	33.9	34.9	29.1	29.4	29.1	29.4
	uniform	28.0	25.5	26.7	24.0	9.6	13.7	33.3	50.0	40.0	28.4	28.4	26.8	30.6
	max.	28.0	25.5	26.7	26.0	31.2	28.4	38.3	100.0	55.4	29.1	33.3	29.1	38.3
resource	ada	26.1	21.2	23.4	23.4	26.4	24.8	35.4	38.3	36.8	28.3	28.6	28.3	29.1
	dt	26.8	22.4	24.4	18.9	18.4	18.6	34.3	40.0	36.9	26.6	26.9	26.7	28.1
	nb	26.8	18.2	21.7	32.0	24.8	27.9	34.1	49.4	40.4	30.9	30.8	30.0	31.9
	gboost	26.7	23.6	25.1	12.8	9.6	11.0	32.2	41.1	36.1	23.9	24.8	24.0	26.6
	knn	30.2	29.1	29.6	33.3	26.4	29.5	39.6	46.7	42.9	34.4	34.1	34.0	35.1
	mlp	31.1	31.5	31.3	24.3	20.8	22.4	36.7	40.0	38.3	30.7	30.8	30.7	31.9
	rf	31.7	26.7	28.9	18.3	16.0	17.1	30.6	37.8	33.8	26.9	26.8	26.6	28.1
	svm	31.2	41.8	35.8	36.0	24.8	29.4	37.4	33.9	35.6	34.9	33.5	33.6	34.3
	average	28.8	26.8	27.5	24.9	20.9	22.6	35.0	40.9	37.6	29.6	29.5	29.2	30.6
interaction	ada	45.2	42.4	43.8	37.0	32.0	34.3	43.0	49.4	46.0	41.7	41.3	41.4	42.3
	dt	40.9	42.4	41.7	38.3	36.8	37.6	48.6	48.3	48.5	42.6	42.5	42.6	43.2
	nb	56.1	38.8	45.9	33.1	72.0	45.3	36.9	17.2	23.5	42.0	42.7	38.2	39.4
	gboost	45.2	42.4	43.8	34.6	29.6	31.9	43.8	50.6	46.9	41.2	40.9	40.9	42.1
	knn	45.5	39.4	42.2	42.6	36.8	39.5	42.9	52.2	47.1	43.7	42.8	42.9	43.6
	mlp	45.5	51.5	48.3	36.5	18.4	24.5	43.6	53.3	48.0	41.9	41.1	40.3	43.4
	rf	56.8	60.6	58.7	46.3	49.6	47.9	54.4	48.3	51.2	52.5	52.8	52.6	53.0
	svm	51.3	46.7	48.9	35.8	30.4	32.9	47.7	56.7	51.8	44.9	44.6	44.5	46.2
	average	48.3	45.5	46.6	38.0	38.2	36.7	45.1	47.0	45.4	43.8	43.6	**42.9**	**44.1**
combined	ada	43.9	41.8	42.9	39.4	32.8	35.8	44.0	51.1	47.3	42.5	41.9	42.0	43.0
	dt	42.8	43.0	42.9	37.8	36.0	36.9	48.1	49.4	48.8	42.9	42.8	42.9	43.6
	nb	56.2	38.2	45.5	33.1	70.4	45.0	38.0	19.4	25.7	42.5	42.7	38.7	39.6
	gboost	44.6	42.4	43.5	35.6	29.6	32.3	43.5	50.6	46.8	41.2	40.9	40.9	42.1
	knn	44.2	39.4	41.7	40.4	35.2	37.6	42.5	50.6	46.2	42.4	41.7	41.8	42.6
	mlp	44.9	53.9	49.0	34.3	19.2	24.6	42.1	47.2	44.5	40.4	40.1	39.4	42.1
	rf	56.6	59.4	58.0	45.7	51.2	48.3	54.1	47.2	50.4	52.2	52.6	52.2	52.6
	svm	50.0	43.0	46.3	33.6	28.8	31.0	45.2	55.6	49.9	43.0	42.5	42.4	44.0
	average	47.9	45.2	46.2	37.5	37.9	36.4	44.7	46.4	45.0	43.4	43.1	**42.5**	**43.7**

$$acc = \frac{TP + TN}{TP + FP + TN + FN}, \tag{13}$$

where TP, FP, TN, and FN denote the number of true positives, false positives, true negatives, and false negatives, respectively. A true positive would mean that a model predicted the same KG class (e.g., high) for a learner as in ground truth.

4.4. Knowledge Gain Prediction

We perform a knowledge gain prediction to assess the influence of the video interaction features on the knowledge gain. We use a recently published evaluation script for the same task [8]. The evaluation script consists of a stratified 10-fold cross-validation for eight classifiers, including hyperparameter optimization and feature selection. To increase the robustness of our results, we repeat the 10-fold cross-validation 5 times with different random states and calculate the metrics based on all predictions.

4.4.1. Classifiers

The classifiers consist of adaboost (ada) [33], decision tree (dt) [34], naive bayes (nb) [35], gradient boosting (gboost) [36], k-nearest neighbors (knn) [37], multilayer perceptron (mlp) [38], random forest (rf) [39], and support vector machine (svm) [40].

4.4.2. Baselines

Additionally to the two feature sets and eight classifiers, we define three baselines that predict (1) the majority class, (2) a stratified distribution, and (3) a uniform distribution based on the training set. These form the lower limit for evaluating the classification results since an equal or lower value would indicate that the features are not better suited as predictors than guessing.

4.4.3. Feature Selection

When dealing with limited data, feature selection is a fundamental preprocessing step. Feature selection is part of hyperparameter optimization; the ideal value is determined based on the validation data in each cross-validation iteration. As the authors in [8], we use the top n features to correlate with the knowledge gain.

4.4.4. Results

The complete classification results are presented in Table 2.

Initially, we observed that video resource features appear to be poor predictors of knowledge gain. None of the classifiers significantly outperforms the baselines, which correspond to (weighted) random guessing. On average, the classifiers achieve a similar F_1-score to the stratified baseline, indicating that these features alone do not provide meaningful predictive power.

In contrast, classification based on interaction features outperforms all baselines. The F_1-score is, on average, $13.7\,\%$ higher than the best baseline, suggesting that user interactions with the videos capture some aspects of learning success. To confirm the robustness of these results, we conducted significance tests on the F_1-score and accuracy across all classifiers based on the baselines for the three different settings, applying Bonferroni correction for multiple comparisons. For a result to be deemed significant, the p-value must satisfy the condition:

$$p < \frac{\alpha}{3} = \frac{0.05}{3} \approx 0.0167, \quad \text{where } \alpha = 0.05. \tag{14}$$

The F_1-score and accuracy for interaction features were significantly better than those of the baselines.

Ultimately, the results for the combined features highlight the positive impact of interaction features on classification performance. This finding is somewhat surprising since the interaction features were calculated independently of the videos' content and design. A plausible hypothesis suggests that interaction features may indirectly capture video content elements. For example, successful users might instinctively pause or rewind during challenging or crucial moments, aligning their actions with the video's complexity or significance. The following experiment will examine which features contributed most significantly to achieving the best classification results.

4.5. Feature Importance

In this experiment, we aim to identify which features contribute most to the best classification example. We use the code provided in Gritz et al. [8] to perform a permutation feature importance analysis. We choose random forest for the interaction features that achieved $52.6\,\%$ accuracy and repeat the evaluation with the corresponding hyperparameters and features from the experiment before. In every iteration of the cross-validation, each feature is discarded 100 times, and the decrease in accuracy is measured. Figure 2 shows the result.

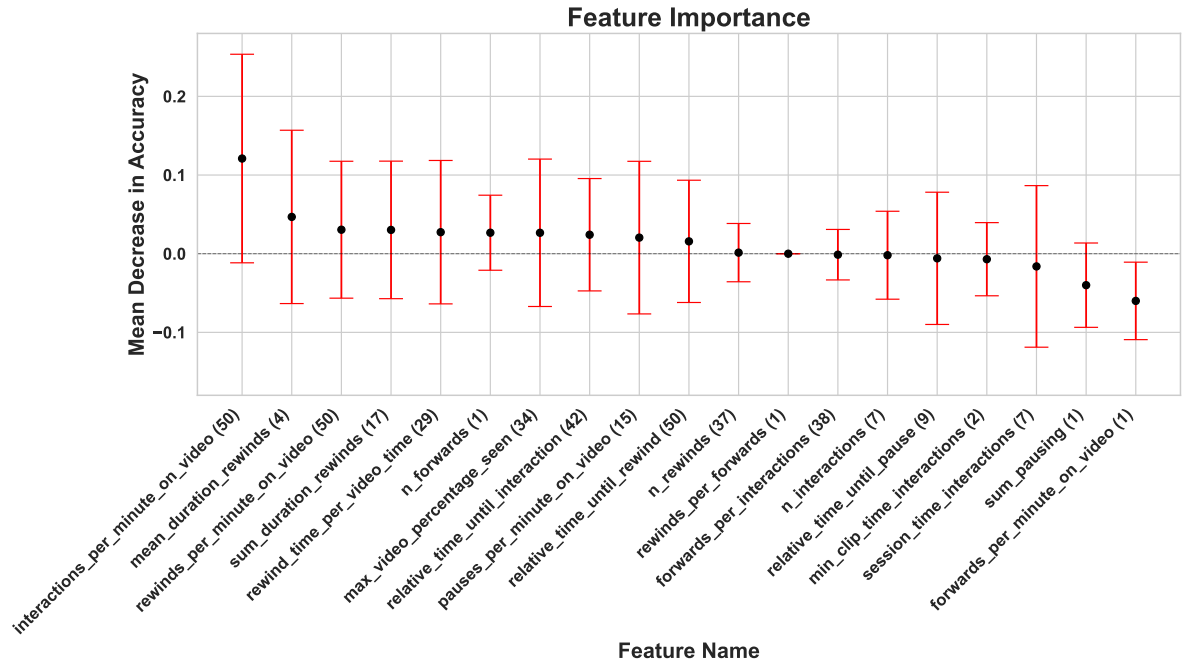

Figure 2: Descending sorted permutation feature importance for the random forest classifier. The number of selections across the cross-validation (max. 50) is displayed on the x-axis in brackets.

The most important feature is the interaction rate, which reflects how actively a learner engages with the videos. Surprisingly, we found a weak negative correlation (Pearson $r = -0.20$) between this feature and the continuous knowledge gain value. The second to fifth important features (although not all selected in every iteration) depend on the rewind interactions. We also found a weak negative correlation (Pearson $r = -0.29$) between the rewind rate and knowledge gain. One possible explanation is that these interactions indicate challenges in learning with videos rather than deeper engagement.

5. Conclusions

In this paper, we investigated the relationship between video interaction features and *knowledge gain* during a learning intended web search on the *SaL-Lightning* dataset [12]. First, we derived interaction logs from screen recordings through a semi-automatic procedure. Next, we developed both (1) features representing the user interactions with the videos seen across the web searches and (2) features indicating the speech rate and (visual) complexity of the video resources. We performed a knowledge gain prediction based on the classification framework provided by recent research [8]. Finally, we analyzed the importance of the individual features in the best classification result.

Surprisingly, the classification results based on the video resource features did not show better values than random guessing. These characteristics may be insufficient to capture the diversity of educational videos and require further research. On the other hand, we observed a $13.7\,\%$ significantly increased F_1-score for the video interaction features, showing that the learning outcomes can be partially explained by the users' interaction with the videos in the learning sessions. Additionally, we found that the interaction rate, especially the rewind rate, is the most important predictor for knowledge gain. These features revealed a weak negative correlation with knowledge gain, indicating that these values might indicate difficulties in learning with videos. However, a limitation of the results is that they were obtained using data from a single study with a single learning task and require verification. Nevertheless, our analysis provides a basis for further research on video-based learning in real-life settings.

These results could be considered when designing assistive tools (e.g., browser add-ons) to support learners actively experiencing difficulties. Furthermore, video designers could adjust their videos accordingly when many users exhibit this behavior (if the information is available). Future research could investigate which aspects of a video lead to increased interaction rates. An additional step could be to predict specific moments in a video that trigger interactions to provide further support (e.g., extra information or system pauses).

Acknowledgments

Part of this work was financially supported by the Leibniz Association, Germany (Leibniz Competition 2023, funding line "Collaborative Excellence", project VideoSRS [K441/2022]).

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly, and DeepL in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style, and Text Translation. After using these tools/services, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Hoppe, P. Holtz, Y. Kammerer, R. Yu, S. Dietze, R. Ewerth, Current challenges for studying search as learning processes, in: Workshop on Learning & Education with Web Data (LILE2018), in conjunction with ACM Web Science, 2018.

[2] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, R. Syed, Assessing learning outcomes in web search: A comparison of tasks and query strategies, in: Conference on Human Information Interaction and Retrieval, CHIIR, Carrboro, USA, ACM, 2016, pp. 163–172. URL: https://doi.org/10.1145/2854946.2854972.

[3] C. Eickhoff, J. Teevan, R. White, S. T. Dumais, Lessons from the journey: a query log analysis of within-session learning, in: International Conference on Web Search and Data Mining, WSDM, New York, NY, USA, 2014, ACM, 2014, pp. 223–232. URL: https://doi.org/10.1145/2556195.2556217.

[4] U. Gadiraju, R. Yu, S. Dietze, P. Holtz, Analyzing knowledge gain of users in informational search sessions on the web, in: C. Shah, N. J. Belkin, K. Byström, J. Huang, F. Scholer (Eds.), Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018, ACM, 2018, pp. 2–11. URL: https://doi.org/10.1145/3176349.3176381. doi:10.1145/3176349.3176381.

[5] W. Gritz, A. Hoppe, R. Ewerth, On the impact of features and classifiers for measuring knowledge gain during web search - A case study, in: Workshops co-located with the International Conference on Information and Knowledge Management, CIKM, Gold Coast, Australia, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-3052/paper6.pdf.

[6] Y. Ghafourian, A. Hanbury, P. Knoth, Readability measures as predictors of understandability and engagement in searching to learn, in: International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, volume 14241 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 173–181. URL: https://doi.org/10.1007/978-3-031-43849-3_15.

[7] R. Yu, R. Tang, M. Rokicki, U. Gadiraju, S. Dietze, Topic-independent modeling of user knowledge in informational search sessions, Information Retrieval Journal 24 (2021) 240–268. URL: https://doi.org/10.1007/s10791-021-09391-7.

[8] W. Gritz, A. Hoppe, R. Ewerth, Unraveling the impact of visual complexity on search as learning, 2025. URL: https://arxiv.org/abs/2501.05289. arXiv:2501.05289.

[9] R. E. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning, 2nd ed., Cambridge University Press, 2014. doi:10.1017/CBO9781139547369.

[10] T.-C. Liu, Y.-C. Lin, S. Kalyuga, Effects of complexity-determined system pausing on learning from multimedia presentations, Australasian Journal of Educational Technology 38 (2021) 102–114. doi:10.14742/ajet.7267.

[11] I. A. Spanjers, T. van Gog, P. Wouters, J. J. van Merriënboer, Explaining the segmentation effect in learning from animations: The role of pausing and temporal cueing, Computers & Education 59 (2012) 274–280. doi:10.1016/j.compedu.2011.12.024.

[12] C. Otto, M. Rokicki, G. Pardi, W. Gritz, D. Hienert, R. Yu, J. von Hoyer, A. Hoppe, S. Dietze, P. Holtz, Y. Kammerer, R. Ewerth, Sal-lightning dataset: Search and eye gaze behavior, resource interactions and knowledge gain during web search, in: D. Elsweiler (Ed.), CHIIR '22: ACM SIGIR Conference on Human Information Interaction and Retrieval, Regensburg, Germany, March 14 - 18, 2022, ACM, 2022, pp. 347–352. URL: https://doi.org/10.1145/3498366.3505835. doi:10.1145/3498366.3505835.

[13] A. Z. Broder, A taxonomy of web search, SIGIR Forum 36 (2002) 3–10. URL: https://doi.org/10.1145/792550.792552.

[14] Y. Ghafourian, A. Hanbury, P. Knoth, Ranking for learning: Studying users' perceptions of relevance, understandability, and engagement, in: International Conference on Theory and Practice of Digital Libraries, TPDL 2023, Zadar, Croatia, volume 14241 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 284–291. URL: https://doi.org/10.1007/978-3-031-43849-3_25.

[15] M. Rokicki, R. Yu, D. Hienert, Learning to rank for knowledge gain, in: Joint Proceedings of the International Workshop on News Recommendation and Analytics (INRA 2022) and the 3rd International Workshop on Investigating Learning During Web Search (IWILDS 2022) co-located with 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, Madrid, Spain, CEUR-WS.org, 2022, pp. 60–68. URL: https://ceur-ws.org/Vol-3411/IWILDS-paper2.pdf.

[16] P. Vakkari, Searching as learning: A systematization based on literature, Journal of Information Science 42 (2016) 7–18. URL: https://doi.org/10.1177/0165551515615833.

[17] N. Roy, F. Moraes, C. Hauff, Exploring users' learning gains within search sessions, in: Conference on Human Information Interaction and Retrieval, CHIIR, Vancouver, Canada, ACM, 2020, pp. 432–436. URL: https://doi.org/10.1145/3343413.3378012.

[18] A. Câmara, D. E. Zein, C. da Costa Pereira, RULK: A framework for representing user knowledge in search-as-learning, in: International Conference on Design of Experimental Search & Information REtrieval Systems, DESIRES, San Jose, USA, CEUR-WS.org, 2022, pp. 1–13. URL: https://ceur-ws.org/Vol-3480/paper-01.pdf.

[19] H. Nasser, D. E. Zein, C. da Costa Pereira, C. Escazut, A. Tettamanzi, RULKKG: estimating user's knowledge gain in search-as-learning using knowledge graphs, in: Conference on Human Information Interaction and Retrieval, CHIIR, Sheffield, United Kingdom, ACM, 2024, pp. 364–369. URL: https://doi.org/10.1145/3627508.3638331.

[20] D. E. Zein, A. Câmara, C. da Costa Pereira, A. Tettamanzi, RULKNE: representing user knowledge state in search-as-learning with named entities, in: Conference on Human Information Interaction and Retrieval, CHIIR, Austin, USA, ACM, 2023, pp. 388–393. URL: https://doi.org/10.1145/3576840.3578330.

[21] D. E. Zein, C. da Costa Pereira, The evolution of user knowledge during search-as-learning sessions: A benchmark and baseline, in: Conference on Human Information Interaction and Retrieval, CHIIR 2023, Austin, TX, USA, ACM, 2023, pp. 454–458. URL: https://doi.org/10.1145/3576840.3578273.

[22] R. Syed, K. Collins-Thompson, Exploring document retrieval features associated with improved short- and long-term vocabulary learning outcomes, in: Conference on Human Information Interaction and Retrieval, CHIIR, New Brunswick, USA, ACM, 2018, pp. 191–200. URL: https://doi.org/10.1145/3176349.3176397.

[23] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, S. Dietze, Predicting user knowledge gain in informational search sessions, in: International Conference on Research & Development in Information Retrieval, SIGIR, Ann Arbor, USA, ACM, 2018, pp. 75–84. URL: https://doi.org/10.1145/3209978.3210064.

[24] C. Otto, R. Yu, G. Pardi, J. von Hoyer, M. Rokicki, A. Hoppe, P. Holtz, Y. Kammerer, S. Dietze, R. Ewerth, Predicting knowledge gain during web search based on multimedia resource consumption, in: International Conference on Artificial Intelligence in Education, AIED, Utrecht, The Netherlands, volume 12748 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 318–330. URL: https://doi.org/10.1007/978-3-030-78292-4_26.

[25] G. D. Rey, M. Beege, S. Nebel, M. Wirzberger, T. H. Schmitt, S. Schneider, A meta-analysis of the segmenting effect, Educational Psychology Review 31 (2019) 389–419. doi:10.1007/s10648-018-9456-4.

[26] M. Merkt, A. Hoppe, G. Bruns, R. Ewerth, M. Huff, Pushing the button: Why do learners pause online videos?, Computers & Education 176 (2022) 104355. doi:10.1016/j.compedu.2021.104355.

[27] C. Dosso, J. G. Moreno, A. Chevalier, L. Tamine, Cost: An annotated data collection for complex search, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4455–4464. URL: https://doi.org/10.1145/3459637.3481998. doi:10.1145/3459637.3481998.

[28] G. Jocher, J. Qiu, Ultralytics yolo11, 2024. URL: https://github.com/ultralytics/ultralytics.

[29] M. Li, T. Lv, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li, F. Wei, Trocr: Transformer-based optical character recognition with pre-trained models, 2021. arXiv:2109.10282.

[30] yt-dlp contributors, yt-dlp, 2025. URL: https://github.com/yt-dlp/yt-dlp, [Computer software].

[31] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[32] A. van Cranenburgh, readability, 2025. URL: https://pypi.org/project/readability/, python package.

[33] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139. URL: https://doi.org/10.1006/jcss.1997.1504.

[34] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[35] T. F. Chan, G. H. Golub, R. J. LeVeque, Updating formulae and a pairwise algorithm for computing sample variances, in: Proceedings of the COMPSTAT Symposium, Springer, Toulouse, France, 1982, pp. 30–41. URL: https://doi.org/10.1007/978-3-642-51461-6_3.

[36] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2001) 1189–1232. URL: http://www.jstor.org/stable/2699986.

[37] E. Fix, J. L. Hodges, Discriminatory analysis - nonparametric discrimination: Consistency properties, International Statistical Review 57 (1989) 238.

[38] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, American Journal of Psychology 76 (1963) 705.

[39] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. URL: https://doi.org/10.1023/A:1010933404324.

[40] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297. URL: https://doi.org/10.1007/BF00994018.