Developing a Multimodal Classroom Engagement Analysis Dashboard for Higher-Education Students

ALPAY SABUNCUOĞLU, UNVEST R&D Center, Turkey
T. METIN SEZGIN, Koc University - Is Bank AI Center, Turkey

Developing learning analytics dashboards (LADs) is a growing research interest as online learning tools have become more accessible in K-12 and higher education settings. This paper reports our *multimodal classroom engagement data analysis* and *dashboard design process* and the resulting engagement dashboard. Our work stems from the importance of monitoring classroom engagement, which refers to students' active physical and cognitive involvement in learning that influences their motivation and success in a given course. To monitor this vital facade of learning, we developed an engagement dashboard using an iterative and user-centered process. We first created a multimodal machine learning model that utilizes face and pose features obtained from recent deep learning models. Then, we created a dashboard where users can view their engagement over time and discover their learning/teaching patterns. Finally, we conducted user studies with undergraduate and graduate-level participants to obtain feedback on our dashboard design. Our paper makes three contributions by (1) presenting a student-centric, open-source *dashboard*, (2) demonstrating *a baseline architecture* for engagement analysis using *our open-access data*, and (3) presenting *user insights and design takeaways* to inspire future LADs. We expect our research to guide the development of tools for novice teacher education, student self-evaluation, and engagement evaluation in crowded classrooms.

CCS Concepts: • **Human-centered computing** → *Interactive systems and tools*; *Visualization systems and tools*; • **Applied computing** → *Education.*

Additional Key Words and Phrases: interactive learning analytics dashboard, classroom engagement, multimodal learning dataset, multimodal data analysis pipeline

ACM Reference Format:
Alpay Sabuncuoğlu and T. Metin Sezgin. 2023. Developing a Multimodal Classroom Engagement Analysis Dashboard for Higher-Education Students. *Proc. ACM Hum.-Comput. Interact.* 7, EICS, Article 188 (June 2023), 23 pages. https://doi.org/10.1145/3593240

1 INTRODUCTION

Classroom engagement refers to students' active involvement in learning, discussion, and reflection with peers, teachers, and materials in a classroom environment. [12]. It is a key factor in predicting the success and future motivation of students in a given course [12]. An experienced teacher can combine several information sources, such as students' affective, behavioral, and cognitive states to examine engagement [46]. Yet, students' self-evaluation skills of classroom engagement are also crucial to analyze their overall learning progress and track their learning habits. Effectively observing individual and group engagement benefits both teachers and students in maintaining a physically and cognitively active classroom environment and motivating both parties to support

Authors' addresses: Alpay Sabuncuoğlu, UNVEST R&D Center, Istanbul, Turkey, asabuncuoglu13@ku.edu.tr; T. Metin Sezgin, Koc University - Is Bank AI Center, Istanbul, Turkey, mtsezgin@ku.edu.tr.

2573-0142/2023/6-ART188
https://doi.org/10.1145/3593240

each other. However, analyzing and interpreting engagement requires grasping multiple components of engagement (affective, behavioral, and cognitive aspects) which complicates students' self-evaluation. Following this concern, reporting the engagement level analysis of students via a dashboard requires critical design decisions.

Learning analytics dashboards (LADs) present information and visualization in a comprehensive summary that curates data related to learning progress in educational contexts [45]. Research demonstrates that LADs have the potential to help inexperienced teachers and students evaluate the data in a meaningful way [11, 48]. Although LADs have gained traction among commercial users and researchers over the past years, it is a still developing field. To explain, as more data sources are becoming available, it requires new analysis methods as well as ways to present this data to users in a comprehensive manner [35, 37]. On the other hand, existing work mostly focuses on the aggregated presentation of numerical and categorical assessment data (e.g.quizzes, real-time tests,video-watching statistics) [35]. Overall, (i) ways to analyze and present multimodal data to users, while (ii) adopting a classroom engagement perspective demands more research. In this respect, in our work, we focused on reporting the multimodal analysis of classroom engagement through a LAD for higher-education students.

Based on the gap outlined above, our LAD design motives were as follows : (1) Presenting an easy-to-follow interface that can accelerate students' self-evaluation of their classroom engagement. (2) Creating a system for students to discover their learning habits and engagement patterns regularly and willingly. (3) Presenting an explainable model to increase transparency. We present what we have learned from users with respect to each motive and the resulting interactive dashboard. Making classroom engagement visible can help students to understand why and where they failed to grasp the concept and how they can succeed following their more engaging moments. Teachers can analyze their classrooms' engagement through different activities and manage their classroom flow accordingly. Considering these possible outcomes, we designed a dashboard that displays the engagement score with a self-assessment test.

Our dashboard contains four main components: (1) Information Area, (2) Engagement Score Chart, (3) Score Prediction Model's Rules, and (4) Cognitive Engagement Quiz Area. The *information area* describes the learning task. *Engagement score* chart visualizes the scores predicted by our baseline multimodal classifier for classroom engagement. We trained this model using our recently open-sourced data that provides audiovisual recordings with students' self-evaluation scores for their engagement levels. We also informed participants about the *score prediction model's rules* to make our model more interpretable. Our model can only predict the engagement scores based on observable features of affective and behavioral engagement such as raising hands and listening carefully. To develop a dashboard that can help students self-evaluate themselves by considering their cognitive engagement, *cognitive engagement quiz area* generates a quiz that automatically assesses students' confidence based on the information area's description.

After the initial wireframe prototype of the dashboard, we dynamically developed the dashboard on Observable through two user studies. Four undergraduate students and three Ph.D. candidates with teaching expertise joined our respective user studies. We inquired about the experience of the dashboard, its perceived benefits, and drawbacks, and obtained participants' suggestions. We analyzed our data in a deductive thematic analysis approach [7] using our dashboard design motives. The insight revealed *four design takeaways* on the reporting format for overall engagement prediction scores, its abstracted visualization, and visual and data-level abstraction consistency

Our research contributes to human-computer interaction and education research with three main outputs:

(1) **Dashboard:** We present a student-centric, open-source, and easy-to-access dashboard that both educators and developers can customize by their needs.
(2) **Dataset and Analysis Pipeline** We demonstrate a baseline architecture for an engagement analysis pipeline using our open-access data, which utilizes state-of-the-art deep learning models in a multimodal format.
(3) **User Insights and Design Takeaways:** We present user insights and four resulting design takeaways for LADs and point at future directions for classroom engagement dashboards.

From a broader perspective, we expect our research to generate tools for student self-evaluation, novice teacher education, and for gauging overall engagement in crowded classrooms.

2 RELATED WORK

This paper focuses on designing a dashboard that effectively reports students' classroom engagement level prediction from a multimodal machine learning model. Throughout our design, development, and study processes, we followed Fredricks et al.'s three-component engagement definition, which comprises affective, behavioral, and cognitive aspects [12]. This section presents an overview of state-of-the-art techniques for analyzing human face and pose, the research in learning analytics dashboards, and the recent developments in explainable AI.

2.1 Deep Face, Pose, and Voice Features

In our model development stage, we combined extracted face and pose features in a tabular form and used these processed information in the training of ML models.

Face Features: Facial expressions are one of the most immediate notifiers of engagement [47]. Recognizing engagement via facial expressions is a long-time research interest in affective computing. One of the most successful and common methods to recognize facial expressions is selecting the action units (AU) for a multistate analysis of face [44]. Tian et al.'s expression analysis system uses forty-four facial action units to describe the facial expressions in a single or additive fashion. Recognizing AUs depends on the model's ability to capture facial representation and find accurate activations for each action unit. Considering the BP4D dataset evaluation on F1 scores, Swin-B-based Graph AU Detection Network [20] and OpenFace [4] promise near state-of-the-art recognition results. We utilized OpenFace's Action Unit Recognition on Multi-Person Videos in our work. OpenFace's Face Feature Extractor yields 1562 features per vector. Our regressor and classifier models use five subsets containing continuous variables: AU Intensities, 3D Eye Landmarks, 3D Face Landmarks, Gaze Directions, and Head Pose.

Pose Features: The human pose estimation task aims to detect the poses of human body parts in 2D or 3D positions. RMPE (AlphaPose) [10], GPN [25], SPM [26], and OpenPose [8] models are some models that demonstrate state-of-the-art recognition results in the MPII-Multi-Person evaluation task [1]. OpenPose is an open-source system for multi-person 2D pose detection in near real-time [8]. The detection results in skeletal joint positions, including body, foot, hand, and facial keypoints. The system also presents near-state-of-the-art results in multi-person pose estimation results [41]. Our work utilizes OpenPose to get insights based on students' actions and interactions with peers.

2.2 Learning Analytics Dashboards

The growing availability of digital educational tools, such as Learning Management Systems (LMSs), Massive Open Online Courses (MOOCs), and other platforms, resulted in a vast amount of educational data [11]. In addition, new multimedia data sources, such as online learning videos and audio recordings, have become available to students and teachers. Parallel to this progress, summarizing, visualizing, and reporting this data has been gaining interest. Learning analytics dashboards (LADs) display this information to different stakeholders in education (students, teachers, governmental agencies, NGOs, etc.) in an effective format with various summarization, and visualization techniques [35]. The aim of a LAD can be showing what has already happened (descriptive), reasoning a particular outcome (diagnostic), predicting the next steps (predictive), or helping the students to achieve their objectives (prescriptive) [48].

Integrating LADs into the classroom and informing teachers with personalized, actionable analytics has also gained some interest in human-AI collaboration [24]. Yet, recent efforts in research and the educational technology market demonstrated that limited technology gain classroom adoption as most teachers lack training in data literacy and developed tools do not fit in the reality of classrooms [15, 42]. So, we should inform both students and teachers through a dashboard with usable, learnable, and trustable insights.

Existing LADs provide analytics either by directly displaying the student data, which requires data literacy, or by making recommendations using blackbox ML models. These models are mainly utilized for predictive outcomes based on students' logged data [32]. Although these blackbox models are biased toward observational data, there is limited research on informing teachers and students about the rationale behind their choice. [38]. To increase the trust in our system, we incorporated recent explainability and interpretability research in our model development stage and integrated these explanations into our dashboard.

2.3 Explainable ML Models

The concept of interpretability in AI focuses on being easily understood and interpreted by humans [5]. Two methods in explainable AI (XAI) research have been gaining weight: Creating glass-box algorithms that can result in transparent rules by definition and creating tools that can explain the behavior of the black-box closed models [14]. Explaining the mechanism of black-box machine learning has no de facto standard. The Explainable AI (XAI) community actively investigates the possible methods for achieving effective communication of explaining and interpreting the machine learning models. One recent method that has achieved widespread adoption is Local Interpretable Model-agnostic Explanations (LIME) [33]. The algorithm can explain the predictions from regressor and classifier models by approximating the local behaviors using pre-defined interpretable models. We utilized the InterpretML toolbox to call the LIME model after the training step to analyze and interpret the explainability rules [27].

3 CLASSROOM ENGAGEMENT AUDIOVISUAL DATA AND MODEL

We developed an engagement score prediction model to present the estimated engagement scores of the students on the dashboard. We trained our model with a subset of our open-sourced dataset [1] [34]. The subset presents two sample learning scenarios with two separate four-student groups. These groups followed Youtube tutorials that requires using computers and completing hands-on activities in approximately one-hour-long group studies. The data contains image sequences of wider angle group recordings that show all students in one frame and individual face sequences. We shared an example frame from these image sequences in Figure 1. In addition, the dataset contains extracted audio recordings and transcripts. The self-evaluation engagement scores of the participants are annotated for each frame and presented in CSV (comma-separated values) format. Participants scored their engagement on a scale between -100 and 100, and these scores were later categorized into five levels, from highly disengaged to highly engaged. Both original scores and processed categorized levels are shared in the dataset.

[1] https://github.com/asabuncuoglu13/classroom-engagement-dataset, *The Classroom Engagement Dataset*

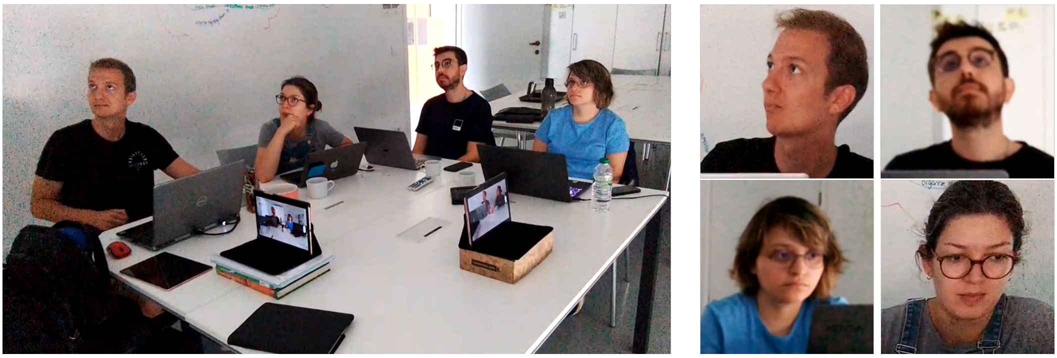


Fig. 1. The left image shows a sample frame from group view recordings. We run the OpenPose on these frames. The four square tiles show individual faces extracted using dlib and FFmpeg frameworks. We run OpenFace on these frames.

3.1 Dataset Format and Statistics

The *Classroom Engagement Dataset* comprises 1280x720 resolution JPEG frames of group views. The individual faces are centered using dlib's feature extractor[2] and cropped to 320x320 resolution. We used FFmpeg for all multimedia operations[3]. For each frame, a self-evaluation-engagement score is available for each participant. The dataset also contains audio recordings in WAV format and transcripts in SRT subtitle format. Figure 1 illustrates one sample frame that shows group view and individual faces.

The first learning session contains 6969 frames. We removed all the low-confidence frames that our feature extractors yielded. In the end, we had 5381 frames usable to train the engagement level classifier for the first group. The second learning session contains 4972 frames. Similarly, we removed all the low-confidence frames, which resulted in 4855 frames usable to train the engagement level classifier for the second group. We used these face and group view frames in the following feature extraction step.

3.2 Feature Extraction

OpenFace: We utilized OpenFace's Face Feature Extractor, which yields 1562 features per vector. We used five continuous feature sets obtained from OpenFace in the training step of our baseline model: AU Intensities, 3D Eye Landmarks, 3D Face Landmarks, Gaze Directions, and Head Pose. We extracted a feature vector for each frame (FPS = 1). The resulting features are interpretable in terms of learning analytics, as the Feature Extractor results in features like eye gaze, head position, action units, etc.

OpenPose: OpenPose's ability to encode global context and part affinity fields (PAF) model's part association has resulted in highly-accurate recognition results for our dataset. We extracted pose features from group videos. OpenPose could extract all four participants' pose features with more than a 0.5 confidence score. For all participants, the resulting feature vector contains twenty-five key points in 3D locations (total length is seventy-five).

[2] http://dlib.net/
[3] https://ffmpeg.org/

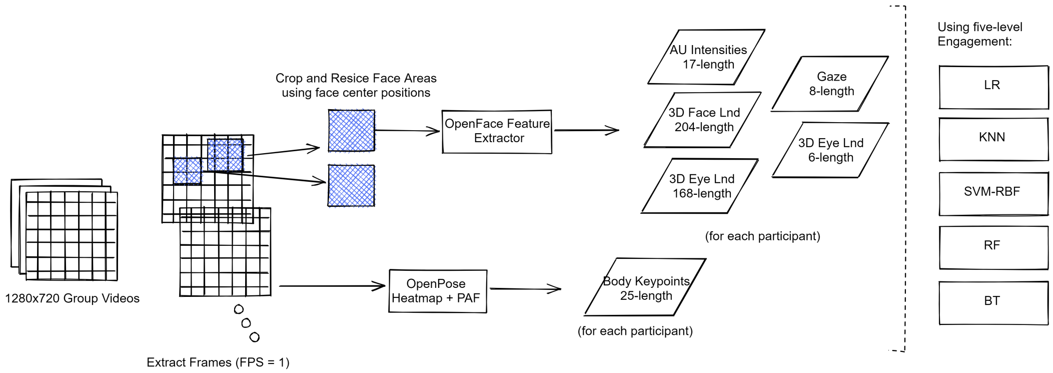


Fig. 2. The illustration of our data pipeline with all components and feature extraction steps. All the code to reproduce the pipeline is shared at https://github.com/asabuncuoglu13/classroom-engagement-dataset.

3.3 Engagement Prediction Model

We tested different machine learning (ML) models that can present simple baseline scores in the dashboard. After utilizing OpenFace to extract facial action units and OpenPose to obtain joint positions of students, we evaluated (*logistic regression* (LR) with *liblinear* solver, *k-Nearest-Neighbor* (KNN), *Support Vector Machine* (SVM) with a *Radial Basis Function* (RBF) kernel, *random forest* (RF) and *boosted trees* (BT)) in classification of five levels of engagement from highly disengaged to engaged. We utilized scikit-learn [28] for all the models' training.

We calculated the weighted F1-score metric with our 80-20 split dataset in the classifier performance evaluation. The F1 score is a harmonic mean of precision and recall. The weighted F1-score calculates scores for each label and finds their weighted average, which provides a more informed decision when data is imbalanced. In the end, the average F1 scores for two models yielded the following results for each algorithm: LR: 0.516, **kNN: 0.684**, RBF-SVM: 0.354, RF: 0.445, and BT: 0.606. Following these results, we used kNN Classifier and Regressor as the baseline model of this study. It was a simple method that presented a baseline for future studies, but it was also adequate to help participants understand their engagement patterns. Figure 2 summarizes our overall data pipeline, which is also available in our open-source repository.

3.4 Interpretability with InterpretML

In this step, we aimed to find the impact of individual features and test the combination of these important features on the same classifiers. We utilized InterpretML [27], an open-source Python library to explain the behavior of existing systems using LIME. A drawback of the current implementation of InterpretML is their limited availability for multi-class classification. So, we produced interpretable rules on binary predictions. We run these algorithms on a binarized version of the data using the classifier. In the end, we shared the rules that have a significant impact on the model's decision in the dashboard. A sample analysis of explainability scores looks like the rules in Table 1.

3.5 Limitations of our Model

3.5.1 Accuracy. Our ML models are trained on a limited dataset in terms of representing the interest levels, learning type variations, and cultural differences. Additionally, the self-evaluation scores compress a complex, three-dimensional engagement definition into a single score. Additionally, existing research demonstrates that the self-evaluation process can yield unreliable results [13]. Future researchers should consider these limitations while adapting our dataset and models.

Component	**Rule**
Data Stats	The binarized engagement level of Session 2 have 4977 *Engaged* observations and 944 *Disengaged* observations.
Highest Score	The highest F1 score with the binarized dataset occurred when we fed 3D Face and Head Pose information, which yielded 0.84.
Action Units	Combination of action units had more impact than individual futures. Combinations with AU5 (upper lid raiser) had more impact than other action units.
Pose Features	All other features could show impact when only combined with other features. The individual features could not show a significant impact (<= 0.1).
Morris Sensitivity [16]	Similarly, Morris sensitivity values only yielded meaningful values for gaze direction and head pose features. In the head pose, we also observed the impact of y-axis values.

Table 1. A sample analysis of LIME feature values for the model developed for Session 2. A different model can yield different important components. So, for different models at different inference times these components might change.

3.5.2 The Maximum number of people. In our classroom deployment scenario, we plan to use 720p cameras to run the engagement level prediction model. The detection of a face is enough to predict an engagement score. As OpenFace requires a minimum of 100px from ear to ear, the model can recognize a maximum of seven faces.

4 DASHBOARD DESIGN

We listed three main motivations to design a classroom engagement dashboard:

(1) Presenting an easy-to-follow interface that can accelerate students' self-evaluation considering affective, behavioral and cognitive levels of engagement. We wanted to provide features that support meaning-making and reflection on the data.
(2) Creating a system for students to discover their learning habits and engagement patterns regularly and willingly.
(3) Increasing the system transparency by providing an explainable model.

In the problem definition, we followed a similar template presented in LATUX (Learning Awareness Tools – User eXperience) [21]. We defined our problem statement following their template for problem definition to standardize our ideation and development process. Table 2 shows the summary of our problem definition with LATUX components.

4.1 Dashboard Components

Considering our problem definition and main motivations, we designed a four-component dashboard. The components are (1) The information area, (2) Engagement Score chart, (3) Rules explanation area, and (4) Cognitive analysis area. We shared the the initial design prototype of the dashboard in Figure 3. In parallel with the findings of previous LAD research, these four areas are designed to present the minimum required information to help students remember the lecture and analyze engagement [22].

The Information Area: This area includes the title and description of the learning task. The teacher is responsible for entering the title and description information in this area.

Problem Def. Component	**Explanation**
Stakeholders	Students, and Teachers. Our current research is not involving other education stakeholders like parents, schools, policy-makers, or NGOs.
Data Sources	Teacher provides a summary description for each learning task. The camera recordings of group activities is fed to our score prediction model, which yields the classroom engagement prediction scores. An external blackbox explainer like LIME yields the interpretability rules.
Data Logging	Currently, logging requires manual labor to run the model and establish the result score CSV. But, the automated process that starts with the camera recording is technologically feasible and possible.
Features of the learning setting	Our research considers only higher-education settings. But, we aim for generalizable results for K-12 education.
Design for Evaluation	From the beginning, we planned to conduct a face-to-face evaluation with small groups.

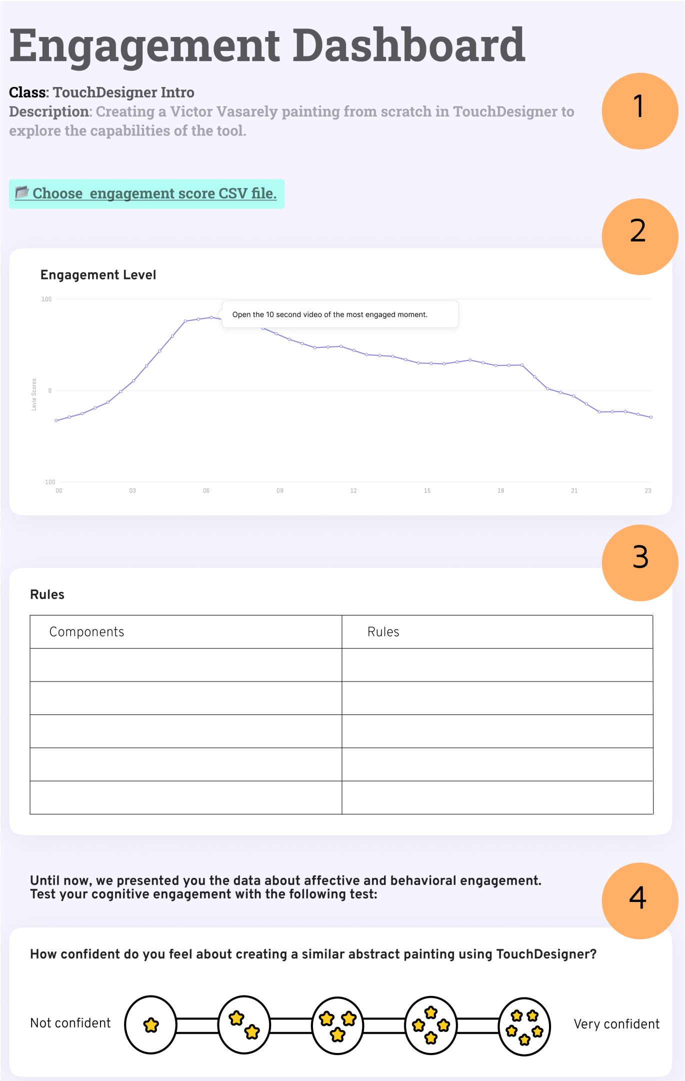
Table 2. The main elements of problem definition.

Engagement Score Chart: This area presents the engagement score obtained from our ML model. It is a frame-based timeline that shows engagement scores for each frame. We also added functionality to display the ten-second video intervals from the video recordings to help students remember their most engaged and disengaged moments. The displayed video includes all group members. In our classroom deployment scenario, the engagement levels are predicted using a single camera that sees all. So, the recordings will come from only one source that includes all group members.

Rules Explanation Area: We aimed to deliver the classification rules determined by LIME in the most readable format. Normally, black box explainers like LIME return the activation points of the models and their relative scores, which are still unreadable for the end user. We included the first fifteen rows of an example LIME result in Appendix C. To make these results more human-readable, we automatically generated more understandable rules in text format. For example, when LIME returns 150 activations with the highest mean absolute score (MAS), like "AU 4 > 0.9234 and AU5 < -0.643 - MAS: 0.3", we report this as "The system determines this score by creating more than a hundred rules. The system could not find one single significant rule. But, when the students have active "Brow Lowerer" and inactive "Upper Lid Raiser" action units, the system is predicted as an "Engaged" moment."

Cognitive Analysis Area: In the multimodal analysis, the machine learning model determines the score based on affective and behavioral engagement checklist items. Determining cognitive engagement requires the self-assessment of the students. We aimed to introduce a small questionnaire in each report to help students assess their cognitive engagement.

The quiz in this area is automatically generated from the description in the Information Area. For example, if the description says, "Creating an abstract painting like Vasarely from scratch in TouchDesigner to explore the capabilities of the tool.", the system can generate a Likert scale question like "What is your level of confidence in creating an abstract painting in TouchDesigner?" or "Mark your confidence level in using different capabilities of the tool." Currently, the system uses a rule-based parser to generate a question from the given nouns. But, it is also possible to fine-tune a language model to generate these cognitive questions from these small descriptions.



Engagement Dashboard

Class: TouchDesigner Intro
Description: *Creating a Victor Vasarely painting from scratch in TouchDesigner to explore the capabilities of the tool.*

📂 Choose engagement score CSV file.

Engagement Level

Open the 10 second video of the most engaged moment.

Rules

Components	Rules

Until now, we presented you the data about affective and behavioral engagement.
Test your cognitive engagement with the following test:

How confident do you feel about creating a similar abstract painting using TouchDesigner?

Not confident ○ ○ ○ ○ ○ Very confident

Fig. 3. The initial design of the dashboard.Four components are numbered on the figure: (1) Information Area, (2) Engagement Score Chart, (3) Rules Explanation Area, (4) Cognitive Analysis Area.

4.2 Software Development

We served the dashboard via Observable [4], an online interactive development environment specifically designed for data-related scenarios. Our main goal in choosing Observable as our development platform was to dynamically change the views and interaction methods in the user studies. The platform choice also supported making the dashboard extendible and modular for researchers, educators, and end-users. The notebooks both present an interactive development environment and a good-looking interface that appeals to the end user. An experienced user with some web-development experience can also tweak the code easily and extend the interface for different use

[4] https://observablehq.com/

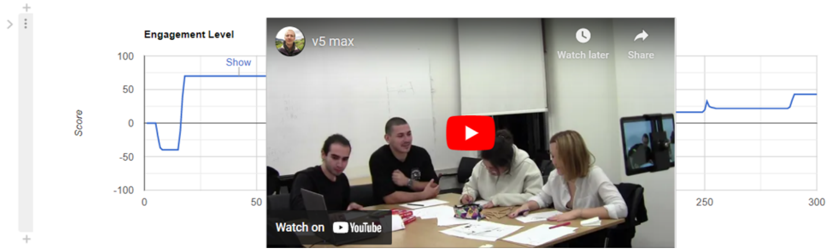


Fig. 4. In Observable, we displayed evey dashboard component in one interactive row. This row displays the Engagement Chart. When the user hovers on Show button, the chart displays corresponding ten-second video.

cases. Figure 4 shows an example row from interactive Observable notebook that displays the *Engagement Chart* with the mouse-hover video display property. All Observable rows shared in Appendix B

5 STUDY AND ANALYSIS

5.1 Participants and Setting

The ethics approval is obtained from the Ethical Committee in Social Research of authors' university. We conducted two respective user studies with the same participants that also attended our *Classroom Engagement Dataset* collection sessions. Both studies took place in the authors' university. The first user study was with four undergraduate students and the second one was with three Ph.D. candidates. The undergraduate students were recruited using University's student forum that explained the study aim and details. The Ph.D. candidates were specifically recruited from the *Interaction Design Lab* as we wanted to gain in-depth insights on the experience and obtain suggestions on the LAD. The Ph.D. participants have completed their grad level lessons and have more than three years of teaching experience. Therefore, the Ph.D. candidates were able to assess both the learning and teaching side of the classroom.

5.2 Study Design

Our dashboard is a one-page interface with four components that consists of novel interactions. To understand the usability and assessment concerns, we conducted a qualitative study that involved talk-aloud sessions rather than a quantitative approach. We prioritized understanding the user's perspective, rather than measuring the statistics of the users' ability to learn and use the system.

In both sessions, we utilized the DATUS (Dashboard Assessment Usability Model) questionnaire as the basis of our discussions. DATUS combines well-known usability metrics with a focus on learning analytics literature, which provided us with a perfect survey to discuss around. The questionnaire also follows the ISO-9241-11 standards and groups its evaluation metrics into eight dimensions: Effectiveness, Efficiency, Satisfaction, Learnability, Accessibility, Recognizability, Aesthetics, and Operability. We shared the questionnaire in Appendix A.

We built our semi-structured discussions inspired by the DATUS questionnaire along these four inquires:

- **(RQ1)** How intuitive is using the dashboard and exploring the engagement data using the provided tools?
- **(RQ2)** How can we develop this dashboard prototype further to meet your expectations?

- **(RQ3)** How can the empirical results inform the design and software updates of the dashboard?

Note that due to the semi-structured nature of the discussions, the questions changed and extended. In addition, we asked additional questions to understand their current practice in learning analytics and asked if and how our dashboard would fit into it.

5.3 Procedure

Our studies roughly followed the following procedure (i) Discussing learning analytics habits, (ii) Introducing Dashboard Cards, (iii) Overall discussion.

In the first session, we met with four undergraduate participants. We showed the initial design prototype of the dashboard (Figure 3) and the participants explored it in a think-aloud fashion. We also asked our semi-structured questions listed earlier. In the meantime, the main author made some updates to the dashboard based on participants' suggestions and concerns in real-time. This way, participants had the chance to observe their ideas in-situ that provided a reflective discussion. In the second session with Ph.D. candidates, we presented the iterated version of the dashboard based on the first user study. Each study session lasted around one hour.

5.4 Measures and Analysis

The study data consisted of audio recordings of semi-structured interviews and the principal researcher's field notes. After familiarizing ourselves with the transcript of the recordings, we thematically grouped the statements with an external researcher [7] in a deductive manner. The motives in the dashboard design guided the themes.

6 RESULTS

Given that our user studies were with two different groups, undergraduate students and Ph.D. candidates, the discussions differed. While the undergraduate students' dominantly talked about the possible benefits and drawbacks of the interface, the Ph.D. candidates lengthily discussed the visualization and possible effects of scoring rules on teachers' and students' behaviors.

Below, we presented our findings in line with our design motivations, by summarizing what we have learned from the user studies to help address our motives more effectively.

6.1 Learning Analytics Habits

We summarized participant's personal analytics habits and listed regularly used analytics interfaces to give an overall view on their data literacy skills.

Physical note-taking: P12 stated that she generally uses her personal notebook and takes notes about task completion. After she shared this habit, all undergraduate participants also noted that they do this kind of note-taking by not realizing it is *a type of learning analytics*. The participants in the second session also use Notion to follow their tasks, but they all use notebooks as a quick tool to check regularly.

Regularly used learning apps: Participant's were not actively using analytics software or checking integrated analytics dashboards. But, they were all familiar with the concept, as they used Blackboard in their classes. Undergrad participant's noted that they only use Blackboard to follow course content and keep track of their grades. They also reported following Coursera, edX, and TreeHouse courses, but they only viewed their general learning progress available on the home page.

Regularly used external analytics apps: As part of our interest, we asked about participants' analytics app use- beyond the learning context. Participants mentioned using ZeppApp (i.e., a sleep pattern app), Youtube Analytics, and screen time reports. All participants emphasized that they like analytics that combines various data sources (e.g., location, duration, etc.) in a comprehensible manner. To illustrate, one participant mentioned that he uses Youtube Analytics regularly, but he strongly dislikes it as it provides many hard-to-read charts all at once. However, they like tracking apps (i.e. fitness or sleeping) as these apps are producing simple, understandable numeric representations and visualizations.

6.2 Providing a Space for Self-Evaluation

Our initial motivation was developing an *easy-to-follow interface that can accelerate students' self-evaluation of their classroom engagement*. Below are the user insights about this point.

Overall score as preferred data format: Seeing a more overall score by combining with other learning data was found more sensible by the participants. In both studies, participants gave examples from sleeping apps. For example, one sleeping tracker app calculates an overall score by looking at all patterns. It makes little sense in the beginning, but when getting used to it, just looking at it can show quick interpretable information for the user.

Visiting the peak moments for interpretation and learning strategy: Participants valued seeing the small sequences of the most and least engaged moments as these would help understand the models' engagement scores. Participants also stated they would visit these moments to see if their engagement was related to the class itself, or if they were distracted/entertained by something else. Therefore, this feature was perceived as a double-check on their score's accuracy. Also, disengaged moments were considered key for exam prep: *"I wish we had something like it in my undergraduate years so that I could just check my disengaged moments and focus on them for the exams."*

Recommendations on boosting one's cognitive engagement: We avoided integrating personalized recommendations in the design of the dashboard. But, undergraduate participants stated that they would like recommendations. *"I would like to see personalized suggestions on how to best follow the courses, similar to sleeping pattern suggestions in sleep tracker apps."*

Valuing individual student engagement: One Ph.D candidate stated that the dashboard would allow her to more effectively evaluate engagement based on each student: *"Mostly, we assess students in a standardized format. Well, we actually do care for individual differences in participation. For example not everyone who talks is cognitively engaged or the silent student could be the one most minds-on with the content. Yet when the classroom size grows, or there's a group work, these individual differences get a bit harder to see... Watching these videos and evaluating them with a machine learning model can enhance the way we interpret student engagement."*

Testing more compact ways to present the data: In the second session, participants asked if we can create a more dense visualization that can immediately show the overall engagement. Using the five-level engagement definition, we can also present the engagement through time in a more compact way as seen in Figure 6. In the studies, we demonstrated this representation on Observable. They stated that, in the compact version, they were looking at the points where the colors change, which is a more visible indicator and eases the evaluation process. P22 stated that, *"... with this representation, I could investigate the red (disengaged) areas in more detail."*

[DT1]: Based on these insights, the first key design takeaway (DT) is **presenting data in an easy-to-grasp format while leaving room for interpretation to provide a space for self-evaluation.**

6.3 Visiting the System Willingly

Our second goal in the dashboard design was *creating a system for students to discover their learning habits and engagement patterns regularly and willingly.*

Using extrinsic motivation to boost intrinsic motivation: All participants noted that only a small percentage of the classroom would self-evaluate themselves. They emphasized that completing the cognitive engagement quiz should be mandatory in the beginning. One participant stated that *"I wouldn't do any quiz if the teacher wouldn't make it mandatory. But, once I start controlling the system regularly and see my positive progress, I would like to visit the dashboard regularly."* This touches upon sources of motivation, a topic widely discussed in education [38]. Participant insights suggest that doing something for an instrumental goal (e.g., grade) known as extrinsic motivation, could initiate intrinsic motivation (e.g., checking the system for its own sake) in the students.

Sending engagement reports in intervals to maintain morale : Participants noted that each class is different and it is not always possible to stay engaged. Therefore, class-by-class engagement reports were not favored. One participant asserted: *"This (disengagement) report would upset me, say if I had a recent break-up... It would show me how I became really disengaged. It could cause stress rather than self-awareness."* This led to a discussion on using intervals (e.g., every four-weeks) for sending engagement reports to students. This would account for the weekly differences.

[DT2]: These insights reveal that, **making use of extrinsic motivators and engagement reports to boost users willingness to check the system.**

6.4 Presenting a Transparent Evaluation

Here we present our findings aligned with our third motive while developing this interface: *Presenting an explainable model to increase transparency.*

Building trust through transparency: All participants reported that they trusted the current model's scoring predictions because it provided the backstage of the results (i.e., the rules). The transparency was valued from a teacher perspective, one Ph.D. candidate noted: *"How should we react when a student comes to us, saying that she is already trying hard, but the system keeps scoring her engagement low? I mean the reasons behind their scores should be clear enough for them to see."* However, the way the rules of the classification system were presented was of critical importance, as we further elaborate in the following items.

The optimal transparency: In the first session with undergrad students, the presentation of rules was not viewed as an important feature of the dashboard. The participants mentioned that seeing the rules is interesting, but they would only look at it once. They suggested that integrating these rule explanations into video can be more engaging. In the second session with Ph.D. candidates, this rule system was discussed more in-depth. These participants noted that the detailed explanations of the model might be exploited by the students. To explain, if the model presented gaze direction as a significant influencer on the score, it would cause students to track teachers' movements and receive high engagement scores. However, it was also noted by participants that this 'mimicking engagement' might have a positive effect since engagement-mimicking behavior can result in real engagement.

Avoiding self-report on cognitive engagement: While presenting rules is a way to introduce the inner workings of the scoring system, the app itself also gives participants a hint. To explain, the cognitive engagement score is obtained from self-evaluation. However, one participant stated that *"I may not trust the cognitive engagement scoring ability of a system by just answering some "how confident" questions. Because maybe I am wrong and I do not actually know my comprehension level."* Another participant noted that answering small and simple questions could be more reliable and would also enhance learning progress.

[DT3]: Offering a transparent evaluation can build trust towards the LAD, but optimal transparency needs to be considered to prevent the users manipulating the scores.

6.5 What to present, when to present

Here, we present our findings on the ideal ways to present the engagement data, that was not a topic we investigated as part of the design motives, but emerged during our analysis of the user insights.

Visibility of analytics' features: While discussing participants' currently used learning analytics apps, they noted that they didn't know the features of platforms. For example, Blackboard integrates many analytics, discussion, and evaluation tools. But the participants did not know this as they cannot easily find these tools. Therefore, it is fundamental to make these analytics visible in the dashboard. While making the analytics visible with reminders or straightforward homepage scores are the first ideas that come to mind, it pertains to some issues discussed below.

Reminders are stressors: The participants didn't like the idea of a regular engagement information reminder. It was asserted as a stressor rather than an informative feature. When we discussed presenting engagement scores not as reminders, but as information on the dashboard, the participants noted similar concerns. Showing the tasks and scores on the home page when they first opened the system was perceived to make them feel stressed. This point is with the following two topics on presentation type that suggests how to combat the perception of data as a 'stressors'.

Avoiding detailed numeric presentations: All participants agreed upon seeing really low values in the Overall Score area would cause disappointment, which would eventually lead to not visiting the interface. For example, if we use our current scoring system between -100 to 100, scoring "0" is moderate engagement, but it could upset the user. So, they stated that they would rather see a score between 0 to 100. So, instead of "0", they would see "50". In addition, some participants could not understand what a negative engagement stands for.

Visualization over metrics: In line with the previous point, participants noted how they preferred abstract visualization rather than scores and metric-based charts. In the second session, participants suggested a visualization system that can illustrate the relative distance to the high engagement. They also stated that this kind of visualization can make the system more playful and would benefit from incorporating gamification.

[DT4]: Presenting an abstract visualization of the engagement scores rather than numeric scores on the homepage to avoid demoralizing the users.

7 REVISITING THE DESIGN AND DISCUSSION

After analyzing the user study results, and listing the possible updates to the dashboard, we revisited the problem definition. Then, we updated the dashboard considering the design takeaways on the reporting format for overall engagement prediction scores, its abstracted visualization, and visual and data-level abstraction consistency.

7.1 Revisiting the Problem Definition

7.1.1 Stakeholders. In the studies, we primarily focused on getting students' perspective, as this dashboard primarily focuses on increasing students' ability to self-evaluate their engagement. Yet, in the second session, comments of Ph.D. candidates also helped us to curate some considerations from a novice teacher perspective. Novice teachers can be one of the main beneficiaries of our dashboard to observe their students and understand their behavioral patterns. Considering our user study results, we defined "teachers" in a more comprehensive way based on their expertise level. In addition, our user study results revealed the students' expectations from teachers come in three-folds: (1) Ensuring everyone interprets the engagement levels clearly, (2) Helping students to evaluate their engagement considering all three aspects of engagement and (3) Moderating the classroom dynamics based on the results of engagement analysis.

7.1.2 Interpretability. In the problem definition, we determined our primary data sources as engagement scores and interpretability rules which were yielded by XAI algorithms like LIME. Our initial aim was to enhance the users' trust in the system. However, combining the dashboard with blackbox model explainers like LIME did not appear as a must-be feature in the studies. In contrast, the students were asking to combine the engagement scores with external personal data sources, such as wearable data, location information and assessment scores, which would result in more complex models with less explainability.

Considering our early results, we can make the deduction that presenting interpretable rules can help students to trust the system only when the presented *rules* are just shallow explanations. As reported earlier, when the rules become more detailed, students might start imitating the behaviors that the model favors. In addition, some concerns over interpretability put extra work on teachers' shoulders. For example, one question raised in the user studies was how teachers should react when a student comes to them, saying that s/he is already trying hard, but the system keeps scoring her/his engagement low. What should be the teacher's reaction? In this case, should students find favor in the camera's eyes or the teacher's eyes? Answering these questions requires a harmonic collaboration between teachers, students and the ML-powered dashboard, where all parties understand the fact that the model just produces some scores based on some patterns, and they should interpret these scores considering their overall engagement progress.

7.1.3 Dealing with motivation. We listed several concerns and challenges in the previous section. A possible unwanted outcome can be causing stress rather than self-awareness. The listed concerns are generalizable for all LADs that show longitudinal data, and should be carefully tracked by educators while using these analytics interfaces. Although limited research is conducted on dashboards' effect on affective states of the students, recent studies by Bennett [6] and Muldner et al. [23] and report by Sclater et al. [36] demonstrated that most students in both K-12 and higher-education levels were motivated by seeing their data presented in a dashboard format, which led to positive behavior changes. Our user studies revealed that the disengaged moments should be carefully treated, as these may cause even more disengagement. It was also noted that disengagement might also stem from non-class related issues (i.e., going through a break-up). Therefore, it was suggested that presenting classroom engagement in intervals (e.g., every four weeks), in ambiguous visualizations over harsh metrics would be more favorable.

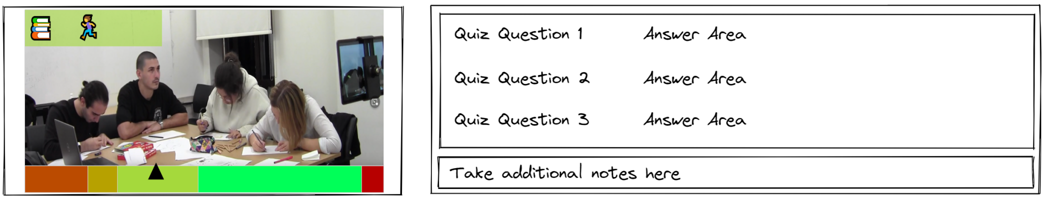


Fig. 5. A sample row of the journal style engagement logging. The left side shows the video with engagement score timeline. The right side asks small auto-generated questions about the course and allow adding additional personal notes.

7.2 Iteration on the LAD

After analysing and interpreting the results of our user studies, we re-designed some components of our dashboard. Figure 5 shows one row of the engagement prototype of the updated version of our interface. Following participants' suggestions, we focused on two main elements while updating the interface: (i) Visualization of class-level and overall engagement, and (ii) consistency of data representation.

(1) *Representing Engagement:* The left side of the row in Figure 6 shows the new stacked bar timeline visualization which displays the categoric engagement levels in a compact form. During the video is playing, a small icon's (running person emoji) distance is changing relative to the engagement icon (books emoji).
(2) *Keeping consistency:* In the previous version, the engagement chart and rules explanation was presenting a very detailed and numeric way of tracking engagement. Compared to the engagement score visualization, cognitive engagement was too shallow, which caused questioning the confidence of the system. In the current version, we included a small quiz related to the activity rather than asking questions about enthusiasm and confidence. Additionally, their overall engagement score can increase by answering these questions correctly, which unified the all interface elements in one representation.

7.2.1 Interactive Rows of a Digital Notebook. We designed the overall view similar to a interactive notebook-like structure. This idea is motivated by their physical notebook usage habits. Following their naturally developed physical notebook habits and suggested features, we designed this format to present the data in a more compact form. Recent research also suggests that integrating classroom data in a format that comes natural to students provides a more exciting interface for students' self-exploration [11, 29]. We also integrated the rules in the overall view, by reminding students about how the model made the decision in the video sections spontaneously.

7.2.2 Engagement Score Abstraction. We created a stacked bar chart to visualize the engagement categories through the lecture. We additionally added an abstract visualization of "engagement distance," where the "Running Human" emoji tries to reach the "Book Stack" emoji, which displays a gamified illustration of the engagement score. Figure 6 shows the new video-watching interface with these abstraction features.

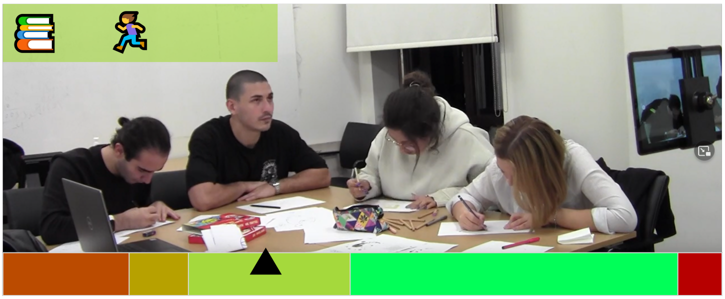


Fig. 6. In the upper-left corner, a colored area (the color is the current engagement category area) shows the distance to the high engagement via emojis. At the bottom, the current video timestamp moves through the engagement category stacked bar chart.

7.2.3 Color Components. We used a traditional gradient from red to green to represent the five-level scale of engagement. Then, we used Viz Palette [19] to check the color space for different types of color blindness. Figure 7 shows the finalized colors and their HEX values in a sample engagement stacked bar timeline chart.

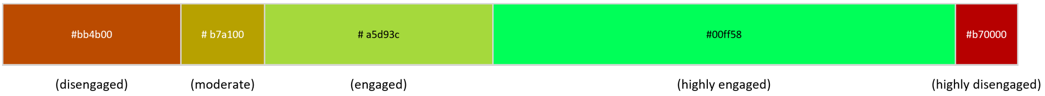


Fig. 7. The timeline view of engagement categories in a stacked bar chart style. From green to red, the HEX values are: ["#b70000", "#bb4b00", "#b7a100", "#a5d93c", "#00ff58"]

7.2.4 Real-time Analysis. Our current ML models do not run in real-time and require extracting external OpenFace and OpenPose features. However, creating a real-time engagement analysis system is technologically feasible, which is a planned future work of this research. Currently, we are using pre-trained OpenFace and OpenPose models to make the system more accurate. However, we can get near real-time results using MediaPipe Face LandMarks [17] and MoveNet [43] models. After collecting more data and developing our own DL model, we can fine-tune and apply quantization techniques to run the final model in-real time.

7.3 Ethical Considerations in Building Our LAD

Our data pipeline and dashboard are designed to aid students and teachers in observing the engagement patterns and interpreting their engagement levels. Yet, one can be concerned about using the dashboard as a classroom surveillance tool, which can process personally-identifiable data that classifies behavior, attitudes, and preferences. Privacy concerns has been a prominent challenge in the learning analytics field [3, 9, 18, 30, 31, 39]. Williamson et al. list four emerging challenges while developing LADs [49]: (1) Protecting participants' privacy while also including enough demographic information, (2) Surveillance concerns, (3) Neglecting pedagogies that fall outside of the dominant narrative and (4) Making LADs maintainable in terms of software development. For each of these emerging issues, we summarize our approach to help researchers and practitioners utilizing our prediction model and dashboard.

(1) **Protecting participant identities:** Our dashboard and model does not require collecting any demographic information. If a third party adapts our model and dashboard, they can run the models anonymously without requiring any identifiers. In the dashboard usage, teachers have access to all data, but students do not have access to other students' engagement levels. Currently, we identified two main users: Teachers and Students. Other stakeholders such as policy-makers might need demographic information to make nationwide decisions. At this point, the engagement data should be aggregated in a privacy-preserving way to protect personal identifiers [2].
(2) **Adressing surveillance concerns:** Our system only give access teachers to personally identifiable data of their students. A student can only see video data from other members of the study group. Our deep learning architectures do not submit any information to third-party software. Our dashboard currently runs on Observable, but the Observable platform does not store any data when the notebook is on run-time. In the dashboard design, we aimed to achieve minimum surveillance concerns. We included a step-by-step explanation of data usage in an end-user-readable way, and we also suggested this approach to adapters of our system.
(3) **Considering implicit pedagogies:** By using our dashboard, students and teachers can explore their learning patterns that fall behind the dominant narrative. The dashboard aims to help students interpret their engagement levels. If other researchers and adapters of the system intend to give suggestions based on their pedagogic approach, they should carefully support their arguments by showing direct links to ML model features.

(4) **Maintaining the software:** Presenting our dashboard on Observable improves the software maintainability significantly. Using Observable, researchers and developers can fork our interactive notebook and create custom dashboards based on their needs. They can also access our data pre-processing scripts and DL model training codes through the project's GitHub repo [5], which is currently active and open-source.

Lastly, we shared these resources with a share-a-like license, so adapters should also make their code open-source to increase maintainability and sustain the ethical considerations.

8 CONCLUSION AND FUTURE WORK

This paper reported our multimodal classroom engagement data analysis and dashboard design process. Our dashboard is specifically designed for students and teachers in higher education. Our research makes three main contributions,

(1) Introducing an online open-source interactive dashboard that both educators and developers can update and extend by their needs. By selecting the available components in the notebook, educators can also conduct their custom A/B tests easily.

(2) Demonstrating an example data pipeline that uses open-access data with easy-to-follow Observable notebooks. We released the code for both classification and interpretability experiments.

(3) Presenting user insights and four design takeaways on LAD design, which gives directions for future classroom engagement dashboards. The design takeaways (DT) are:
[DT1] Presenting data in an easy-to-grasp format while leaving room for interpretation to provide a space for self-evaluation.
[DT2] Making use of extrinsic motivators and engagement reports to boost users willingness to check the system.
[DT3] Offering a transparent evaluation can build trust towards the LAD, but optimal transparency needs to be considered to prevent the users manipulating the scores.
[DT4] Presenting an abstract visualization of the engagement scores rather than numeric scores on the homepage to avoid demoralizing the users.

Our LAD design process and the learnings can benefit researchers from all disciplines that would like to design a learning analytics dashboard. From a broader perspective, we expect our research to be integrated into novice teacher education by providing an easy-to-use tool to give hints about groups' engagements and help them build engagement-related conversations with students. We expect our models and dashboards to be a part of crowded classrooms as a helper tool to better evaluate overall engagement.

ACKNOWLEDGMENTS

This project is supported by Koc University - Is Bank AI Center. We would like to thank for their genereous support. We would like to thank all participants to become a part of the dataset and our research.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2D Human Pose Estimation: New Benchmark and State of the Art Analysis. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3686–3693. https://doi.org/10.1109/CVPR.2014.471

[2] Benny Applebaum, Haakon Ringberg, Michael J. Freedman, Matthew Caesar, and Jennifer Rexford. 2010. Collaborative, Privacy-Preserving Data Aggregation at Scale. In *Privacy Enhancing Technologies*, Mikhail J. Atallah and Nicholas J. Hopper (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 56–74.

[5] https://github.com/asabuncuoglu13/classroom-engagement-dataset

[3] Kimberly E. Arnold and Niall Sclater. 2017. Student perceptions of their privacy in leaning analytics applications. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 66–69. https://doi.org/10.1145/3027385.3027392

[4] Tadas Baltrusaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6. https://doi.org/10.1109/FG.2015.7284869

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[6] Liz Bennett. 2018. *Students' learning responses to receiving dashboard data: Research Report*. Technical Report. https://doi.org/10.13140/RG.2.2.18504.83205

[7] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. https://doi.org/10.1080/2159676X.2019.1628806 arXiv:https://doi.org/10.1080/2159676X.2019.1628806

[8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2019. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. arXiv:1812.08008 [cs] http://arxiv.org/abs/1812.08008

[9] Hendrik Drachsler and Wolfgang Greller. 2016. Privacy and analytics: it's a DELICATE issue a checklist for trusted learning analytics. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge - LAK '16*. ACM Press, Edinburgh, United Kingdom, 89–98. https://doi.org/10.1145/2883851.2883893

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2018. RMPE: Regional Multi-person Pose Estimation. arXiv:1612.00137 [cs] http://arxiv.org/abs/1612.00137 version: 5.

[11] Gloria Milena Fernandez Nieto, Kirsty Kitto, Simon Buckingham Shum, and Roberto Martinez-Maldonado. 2022. Beyond the Learning Analytics Dashboard: Alternative Ways to Communicate Student Data Insights Combining Visualisation, Narrative and Storytelling. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, 219–229. https://doi.org/10.1145/3506860.3506895

[12] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School Engagement: Potential of the Concept, State of the Evidence. *Review of Educational Research* 74, 1 (March 2004), 59–109. https://doi.org/10.3102/00346543074001059

[13] Nan Gao, Mohammad Saiedur Rahaman, Wei Shao, and Flora D Salim. 2021. Investigating the Reliability of Self-Report Data in the Wild: The Quest for Ground Truth. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers (UbiComp '21)*. Association for Computing Machinery, New York, NY, USA, 237–242. https://doi.org/10.1145/3460418.3479338

[14] Hani Hagras. 2018. Toward Human-Understandable, Explainable AI. *Computer* 51, 9 (2018), 28–36. https://doi.org/10.1109/MC.2018.3620965

[15] Sarah K. Howard. 2013. Risk-aversion: understanding teachers' resistance to technology integration. *Technology, Pedagogy and Education* 22, 3 (2013), 357–372. https://doi.org/10.1080/1475939X.2013.802995

[16] Bertrand Iooss and Andrea Saltelli. 2017. *Introduction to Sensitivity Analysis*. Springer International Publishing. 1103–1122 pages. https://doi.org/10.1007/978-3-319-12385-1_31

[17] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. https://doi.org/10.48550/ARXIV.1907.06724

[18] Charles Lang, Charlotte Woo, and Jeanne Sinclair. 2020. Quantifying data sensitivity: precise demonstration of care when building student prediction models. In *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*. ACM, Frankfurt Germany, 655–664. https://doi.org/10.1145/3375462.3375506

[19] Susie Lu and Elijah Meeks. 2022. Viz Palette. https://projects.susielu.com/viz-palette

[20] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. 2022. Learning Multi-dimensional Edge Feature-based AU Relation Graph for Facial Action Unit Recognition. (July 2022), 1239–1246. https://doi.org/10.24963/ijcai.2022/173

[21] Roberto Martinez-Maldonado, Abelardo Pardo, Negin Mirriahi, Kalina Yacef, Judy Kay, and Andrew Clayphan. 2015. The LATUX workflow: designing and deploying awareness tools in technology-enabled learning settings. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, 1–10. https://doi.org/10.1145/2723576.2723583

[22] Wannisa Matcha, Nora'ayu Ahmad Uzir, Dragan Gašević, and Abelardo Pardo. 2020. A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies* 13, 2 (2020), 226–245. https://doi.org/10.1109/TLT.2019.2916802

[23] Kasia Muldner, Michael Wixon, Dovan Rai, Winslow Burleson, Beverly Woolf, and Ivon Arroyo. 2015. Exploring the Impact of a Learning Dashboard on Student Affect, Cristina Conati, Neil Heffernan, Antonija Mitrovic, and M. Felisa

Verdejo (Eds.). Springer International Publishing, Cham, 307–317.

[24] Tanya Nazaretsky, Carmel Bar, Michal Walter, and Giora Alexandron. 2018. Empowering Teachers with AI: Co-Designing a Learning Analytics Tool for Personalized Instruction in the Science Classroom. (2018), 16.

[25] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. 2017. Generative Partition Networks for Multi-Person Pose Estimation. arXiv:1705.07422 [cs] http://arxiv.org/abs/1705.07422 version: 2.

[26] Xuecheng Nie, Jianfeng Zhang, Shuicheng Yan, and Jiashi Feng. 2019. Single-Stage Multi-Person Pose Machines. arXiv:1908.09220 [cs] http://arxiv.org/abs/1908.09220 version: 1.

[27] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. arXiv:1909.09223 [cs, stat] http://arxiv.org/abs/1909.09223

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

[29] Stanislav Pozdniakov, Roberto Martinez-Maldonado, Yi-Shan Tsai, Mutlu Cukurova, Tom Bartindale, Peter Chen, Harrison Marshall, Dan Richardson, and Dragan Gasevic. 2022. The Question-driven Dashboard: How Can We Design Analytics Interfaces Aligned to Teachers' Inquiry?. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, 175–185. https://doi.org/10.1145/3506860.3506885

[30] Paul Prinsloo and Sharon Slade. 2015. Student privacy self-management: implications for learning analytics. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*. ACM, Poughkeepsie New York, 83–92. https://doi.org/10.1145/2723576.2723585

[31] Paul Prinsloo and Sharon Slade. 2017. An elephant in the learning analytics room: the obligation to act. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. ACM, Vancouver British Columbia Canada, 46–55. https://doi.org/10.1145/3027385.3027406

[32] Irina Rets, Christothea Herodotou, Vaclav Bayer, Martin Hlosta, and Bart Rienties. 2021. Exploring critical factors of the perceived usefulness of a learning analytics dashboard for distance university students. *International Journal of Educational Technology in Higher Education* 18, 1 (2021), 46. https://doi.org/10.1186/s41239-021-00284-9

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs, stat] http://arxiv.org/abs/1602.04938

[34] Alpay Sabuncuoglu and T. Metin Sezgin. 2023. Multimodal Group Activity Dataset for Classroom Engagement Level Prediction. arXiv:cs.HC/2304.08901

[35] Beat A. Schwendimann, Maria Jesus Rodriguez-Triana, Andrii Vozniuk, Luis P. Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. 2017. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies* 10, 1 (2017), 30–41. https://doi.org/10.1109/TLT.2016.2599522

[36] Niall Sclater, Alice Peasgood, and Joel Mullan. 2016. *Learning Analytics in Higher Education: A review of UK and international practice*. Technical Report. https://www.jisc.ac.uk/sites/default/files/learning-analytics-in-he-v2_0.pdf

[37] Gayane Sedrakyan, Erik Mannens, and Katrien Verbert. 2019. Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages* 50 (2019), 19–38. https://doi.org/10.1016/j.jvlc.2018.11.002

[38] Shiva Shabaninejad, Hassan Khosravi, Solmaz Abdi, Marta Indulska, and Shazia Sadiq. 2022. Incorporating Explainable Learning Analytics to Assist Educators with Identifying Students in Need of Attention. In *Proceedings of the Ninth ACM Conference on Learning @ Scale*. ACM, 384–388. https://doi.org/10.1145/3491140.3528292

[39] Sharon Slade, Paul Prinsloo, and Mohammad Khalil. 2019. Learning analytics at the intersections of student trust, disclosure and benefit. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*. ACM, Tempe AZ USA, 235–244. https://doi.org/10.1145/3303772.3303796

[40] Rita Sofia da Silva Antunes. 2020. *DATUS: Dashboard Assessment Usability Model: A case study with student dashboards*. mastersthesis.

[41] Liangchen Song, Gang Yu, Junsong Yuan, and Zicheng Liu. 2021. Human pose estimation and its application to action recognition: A survey. *Journal of Visual Communication and Image Representation* 76 (2021), 103055. https://doi.org/10.1016/j.jvcir.2021.103055

[42] James P. Spillane. 2012. Data in Practice: Conceptualizing the Data-Based Decision-Making Phenomena. *American Journal of Education* 118, 2 (2012), 113–141. https://doi.org/10.1086/663283

[43] Tensorflow. 2023. movenet/multipose/lightning. https://tfhub.dev/google/movenet/multipose/lightning/1

[44] Ying-li Tian, Takeo Kanade, and Jeffrey F Cohn. 2001. Recognizing Action Units for Facial Expression Analysis. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE* 23, 2 (2001), 19.

[45] Katrien Verbert, Sten Govaerts, Erik Duval, Jose Luis Santos, Frans Van Assche, Gonzalo Parra, and Joris Klerkx. 2014. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing* 18, 6 (Aug. 2014), 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

[46] Ze Wang, Christi Bergin, and David A. Bergin. 2014. Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly* 29, 4 (2014), 517–535. https://doi.org/10.1037/spq0000050

[47] Jacob Whitehill, Zewelanji Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. 2014. The Faces of Engagement: Automatic Recognition of Student Engagementfrom Facial Expressions. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING* 5, 1 (2014), 13.

[48] Kimberly Williamson and Rene Kizilcec. 2022. A Review of Learning Analytics Dashboard Research in Higher Education: Implications for Justice, Equity, Diversity, and Inclusion. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, 260–270. https://doi.org/10.1145/3506860.3506900

[49] Kimberly Williamson and Rene Kizilcec. 2022. A Review of Learning Analytics Dashboard Research in Higher Education: Implications for Justice, Equity, Diversity, and Inclusion. In *LAK22: 12th International Learning Analytics and Knowledge Conference*. ACM, Online USA, 260–270. https://doi.org/10.1145/3506860.3506900

A DATUS QUESTIONNAIRE

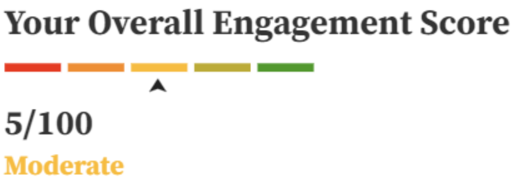
The below survey items are answered with a 5-level Likert scale. (Strongly Disagree to Strongly Agree) [40]

(1) Overall, I am satisfied with how easy it is to use this dashboard.
(2) It was simple to use this dashboard.
(3) I can effectively complete my work using this dashboard.
(4) I am able to complete my goals (tasks) quickly using this dashboard.
(5) I am able to efficiently complete my goals (tasks) using this dashboard.
(6) I feel comfortable using this dashboard.
(7) It was easy to learn to use this dashboard.
(8) I believe I became productive quickly using this dashboard.
(9) Whenever I make a mistake using the dashboard, I recover easily and quickly.
(10) The information (on-screen messages) provided with this dashboard is clear.
(11) It was easy to find the information I needed.
(12) The information displayed in the dashboard is easy to understand.
(13) The information displayed in the dashboard is effective in helping me complete the tasks and scenarios.
(14) The organization of the information on the dashboard is clear.
(15) The interface of this dashboard is pleasant.
(16) I like using the interface of this dashboard.
(17) This dashboard has all the functions and capabilities I expect it to have.
(18) Overall, I am satisfied with this dashboard.
(19) Data on the dashboard is easy to read.
(20) Visual encoding of data is consistent throughout the dashboard.

B DASHBOARD PARTS IN OBSERVABLE

We developed our Dashboard in Observable, an interactive notebook platform for data-powered applications. The notebook is available at https://observablehq.com/d/21637c01efc12a65.

B.1 Displaying Overall Engagement



Your Overall Engagement Score

5/100

Moderate

Fig. 8. We displayed the overall engagement using a caret pointing on a discrete gradient bar. The chart implementation is an updated version of @duckyb's fever-chart

B.2 Engagement Chart

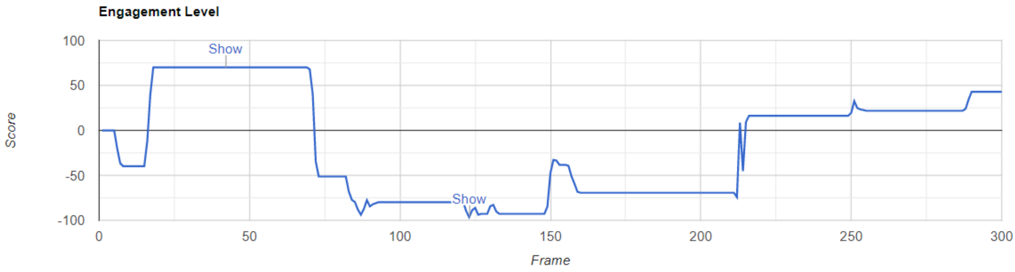
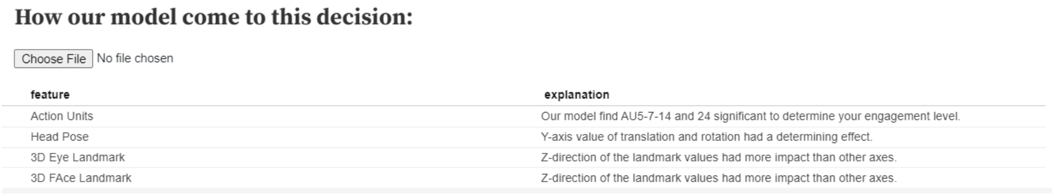


Fig. 9. We used Google Charts to show display the engagement scores through time. When the user hovers on Show button, the chart displays corresponding ten-second video.

B.3 Rules Explanations

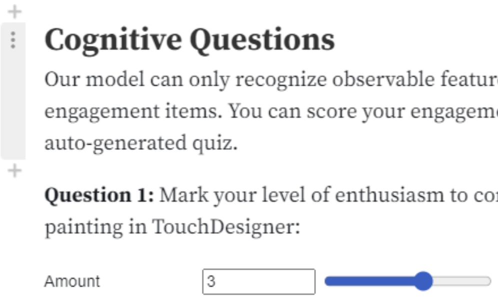


How our model come to this decision:

feature	explanation
Action Units	Our model find AU5-7-14 and 24 significant to determine your engagement level.
Head Pose	Y-axis value of translation and rotation had a determining effect.
3D Eye Landmark	Z-direction of the landmark values had more impact than other axes.
3D FAce Landmark	Z-direction of the landmark values had more impact than other axes.

Fig. 10. Rules are extracted from the LIME algorithm and later simplified manually in the produced CSV. In the Observable implementation, we manually selected the rules file and displayed it on tabular format.

B.4 Cognitive Analysis Test



Cognitive Questions

Our model can only recognize observable features of *affective* and *behavioral* engagement items. You can score your engagement level by taking the following auto-generated quiz.

Question 1: Mark your level of enthusiasm to continue creating another abstract painting in TouchDesigner:

Amount 3

Fig. 11. We showed two questions about their enthusiasm and confidence, based on the description of the current learning task. Participants used a slider to select their level between 0-5.

C EXAMPLE OUTPUT OF LIME

We used InterpretML to explain feature importance of our blackbox classifiers. In the *Rules Explanation Area* of our dashboard, we presented a simplified version of the InterpretML's feature importance output. The following output (Table 3) is the first 15 lines of the 500-line-long output. The output shows the SVM model's explanation. The all raw outputs are accessible at https://github.com/asabuncuoglu13/classroom-engagement-dataset/tree/main/eda/classifiers/reports

Rule, Type, Coef, Support, Importance
346,pose_Rx <= 1.548622965812683,rule,0.1710477137102923,0.970703125,0.028845039401672184
1356,Z_20 <= 2.5053281784057617 & AU04_r <= 2.580228090286255 & AU12_r <= 3.9592233896255493,rule,0.004212891108200999,0.962890625,0.0007963620081684719
856,AU04_r <= 3.5617398023605347 & Z_22 <= 2.0053879022598267 & eye_lmk_X_15 > -2.6626787185668945,rule,0.14634273358518968,0.947265625,0.03270799732117178
618,gaze_1_y > -1.569492220878601,rule,-0.07226943523147979,0.943359375,0.016705414815918324
1188,Y_0 <= 1.555420160293579,rule,-0.020016843049618795,0.9296875,0.00511776511575324
1535,Y_2 <= 1.5189287662506104,rule,-0.06285284564964531,0.923828125,0.016673157703120788
1706,pose_Rx <= 1.8917128443717957 & AU14_r <= 1.5858372449874878,rule,-0.07540847056964113,0.921875,0.020237234612891498
1670,Y_14 <= 1.4680776596069336,rule,-0.05259229652494245,0.921875,0.014114099325527318
436,Y_15 <= 1.4759466052055359,rule,-0.05249128261956105,0.912109375,0.01486215961002277
1283,pose_Tz > -1.0056792497634888,rule,0.028965973666379406,0.912109375,0.008201303195609496
1641,Y_21 <= 1.4215712547302246 & Y_24 <= 1.4029836654663086 & eye_lmk_X_3 > -2.5689715147018433,rule,0.03797498576264193,0.900390625,0.011372690217192762
1778,Y_14 <= 1.331681251525879,rule,-0.0406223509960535,0.900390625,0.01216551907773639
591,Y_17 <= 1.371969223022461 & Y_26 <= 2.3223878145217896,rule,0.06639462533729454,0.89648437
1541,Z_60 <= 1.2910636067390442,rule,-0.06219444549003456,0.88671875,0.019711676466791
897,gaze_0_z > -1.148589849472046,rule,-0.06382887839723846,0.88671875,0.02022968755957648

Table 3. Raw output of the first fifteen lines of IntepretML's LIME explanation.

Received July 2022; revised October 2022; accepted December 2022.