Article

Predictive Analytics for Student Engagement in E-Learning Systems

Yerkebulan Murattaly [1] and Azamat Serek* [2]

[1]Department of Computer Science, SDU University, Kaskelen, Kazakhstan
[2]Astana IT University, Astana, Kazakhstan

DOI: 10.47344/88c0mc52

Abstract

To increase the success of students' education, it is important to be able to predict the level of their involvement in the online educational environment. This study uses the Open University Learning Analytics (OULAD) open dataset to develop a systematic and reproducible approach to classifying student engagement. On the other hand, many other studies depend on specific datasets or limited definitions of engagement. A full cycle of data preprocessing and feature extraction was implemented, aimed at obtaining informative behavioral indicators based on click data and evaluation results. We trained and tested two traditional supervised machine learning model, Random Forest and Logistic Regression, using weight and macro-average metrics. The random forest model demonstrated high efficiency across all interaction classes and showed higher accuracy (0.926) compared to logistic regression (0.896). The results obtained emphasize the importance of high-quality data preprocessing and thoughtful design of features. In addition, they confirm that such signs provide valuable information for the development of early warning systems and the further development of educational analytics in higher education institutions.

Keywords: student engagement, learning analytics, feature engineering, preprocessing pipelines, classical machine learning, Logistic Regression, Random Forest, OULAD

I. Introduction

The rapid development of online and high-tech educational environments has significantly changed modern education. Learning management systems and virtual learning environments, widely used in distance learning, is now an integral element of both fully online and blended learning models [14]. With the growing popularity of such platforms, it is becoming especially important for educational institutions to understand and monitor the level of student engagement, as well as to monitor academic performance, re-education, and course completion [15], [16]. Student engagement, usually determined by the level of attention, activity, and effort expended in the learning process, is widely considered in scientific research as a reliable indicator of the effectiveness of the educational process [3], [17], [18].

Email (1): 221107028@stu.sdu.edu.kz ORCID: 0009-0006-1210-6015
*Corresponding author:
Email: Azamat.Serek@astanait.edu.kz ORCID: 0000-0001-7096-6765

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

Modern advances in learning analytics have made it possible to systematically analyze student interaction logs, assessment data, and behavioral patterns recorded in virtual learning environments (VLE). These datasets offer numerous opportunities for building predictive models, but because they are multidimensional, sparse, and heterogeneous, they require preprocessing to create useful features. As a result of high-quality processing of the source data from the LMS system, they are transformed into important features that increase the interpretability of models, predictive accuracy and reproducibility of results. As a result, the reliability of predicting the level of involvement of students increases. [2], [4].

An open data collection called Open University Learning Analytics (OULAD) was used to stimulate work in the field of research. This data set contains demographic data, assessment results and detailed log files reflecting their activities in the educational process in relation to 32,593 students who have completed 22 training courses. The total amount of data is more than 10.6 million activity records per day. Such a large and informative data set makes it particularly effective for forecasting tasks.g [1].

In this study, we examined five supervised machine learning algorithms—random forest (RF), gradient boosting (GB), AdaBoost, logistic regression (LR), and support vector machines (SVM) to classify student engagement levels. For a more in-depth study, we chose logistic regression and random forest because they are highly accurate, stable, and easy to understand. This choice allowed us to systematically compare the models, highlighting the importance of conducting accurate comparative analysis and obtaining consistent results.

This paper presents three main scientific contributions: (1) developing an integrated feature engineering and data preprocessing methodology for engagement prediction using OULAD behavioral data; (2) evaluating the performance of two well-known supervised learning models: random forest and logistic regression; and (3) empirically demonstrating the most important features and algorithms for engagement classification. By focusing on data preprocessing procedures and using traditional machine learning models, the results provide a replicable framework and valuable recommendations for developing early warning systems and improving educational analytics in higher education institutions.

II. Literature Review

Recently, researchers have paid special attention to student engagement in online learning, as it is directly related to academic achievement, student retention, and course completion [15], [17]. Despite the growing popularity of this topic, many unanswered questions remain regarding how to measure engagement, how to select factors for evaluation, and how to make accurate predictions based on log data from educational platforms [16]. In addition to summarizing the findings of previous studies, this review highlights the crucial role of data preprocessing and feature generation in predicting student engagement.

A. Definitions and Aspects of Engagement in E-Learning

Digital learning engagement among students is a complicated and a multidimensional issue. There are various opinions in terms of the effect on digital learning in education. There can be various possibilities what will be results of such kind of education. The majority of learning analytics research in higher education is based on observable behavioral indicators such clicks, task completion time, and login frequency, whereas the cognitive, social, and emotional aspects of engagement are frequently overlooked. Observable features are used because it can be easier to get them and analyse thereafter, Although log data provides a more convenient way to quantify behavioral engagement, this constraint restricts the efficacy of engagement models and the depth of understanding of student participation [16].

Diversity and ambiguity in terminology, measurement methods, and annotation criteria across multiple studies and datasets have been identified in recent reviews of automated engagement assessment [8], [9], [16]. Currently diversification is undergoing increase. Only a few of them correlate with validated psychological scales, which makes it difficult for cross-study comparisons and generalization of results [15], [17]. In this regard, the importance of using frameworks that cover the behavioral, cognitive, social, and emotional aspects of engagement is emphasized, as well as the need for careful data preprocessing to transform heterogeneous log data into informative signs of interaction [17]. These frameworks highlight the importance of preprocessing: changing mixed raw data into clear and meaningful features that capture different parts of the engagement. Preprocessing is also very important in terms of subsequent machine learning application because the data should be clean before we actually apply machine learning.

B. Practical Application of Learning Analytics for Engagement or Performance Forecasting

Research based on online learning data aims to predict the level of engagement and learning outcomes of students. Early work has shown that predictive models that take into account behavioral, collaborative, and emotional components can serve as an early

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

warning system [5]. Other studies focus on analyzing behavioral interactions in self-regulated and survey-oriented online modules, including interaction with content, duration of activity, and resources used, demonstrating a close relationship between behavioral traits and learning outcomes and engagement levels.

The consistency and sustainability of the engagement metrics extracted from LMS journals have a huge impact on academic performance, and machine learning models trained on such data that usually demonstrate high predictive accuracy [16]. Developing multimodal approaches by combining the analysis of gaze, facial expressions and actions shows that the integration of visual and behavioral data makes it possible to more accurately assess engagement compared to using a single source of information [17]. At the same time, the majority of studies emphasize that the quality of data preprocessing and feature development directly determines the accuracy of forecasting and interpretability of models.

C. Challenges, Limitations, and Gaps in Existing Research

Despite the progress made, there are still a number of challenges in this area. Many studies focus primarily on behavioral indicators, while the cognitive and emotional aspects of the engagement remain insufficiently considered [5], [15]. Approaches to defining interaction protocols and annotations are inconsistent, and the use of validated scales is found only in a limited number of papers [8], [16], [17]. Large publicly available datasets are used relatively rarely, which reduces the possibility of generalizing the results and reproducibility of research. Comparative studies of classical supervised machine learning algorithms are also insufficient, which makes it difficult to identify sustainably effective methods. In addition, works devoted to replication and comparative analysis are rare due to the variety of data sets and methodological approaches used, and the stages of data preprocessing are often described superficially, which limits the reproducibility and practical applicability of engagement forecasting models.

D. Implications for the Current Study and Knowledge Deficiency

The literature emphasizes that predicting engagement is a dynamically developing field, but most research is limited to rigid definitions, inconsistent indicators, and the use of specialized or small datasets. In this regard, there is an obvious need for systematic research based on large publicly available data, the use of supervised machine learning methods and clearly described procedures for preprocessing and feature development [8].

Using the Open University Learning Analytics (OULAD) dataset and extracting meaningful behavioral characteristics through a rigorous data preparation cycle, this work fills these gaps. Due to their high predictive performance, interpretability, and robustness, logistic regression and random forest were selected for further investigation. By highlighting the critical role of data preprocessing in obtaining consistent and understandable results, the proposed methodology establishes a replicable standard and provides empirical evidence for the effectiveness of traditional engagement prediction algorithms [5].

III. Methods

This section showcases the methods that we used and the methodology to achieve research aim that we outlined in the introduction section.

A. Dataset Description and Rationale

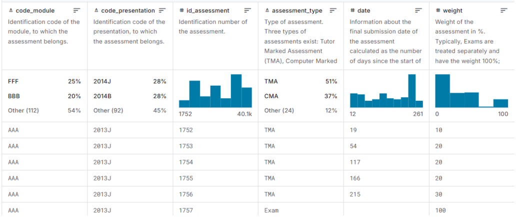
This work utilized open source dataset. The study used the Open University Learning Analytics (OULAD) dataset, which includes complete log-records describing student demographic characteristics, course metadata, evaluation metrics, and virtual learning environment (VLE) activities [21].

Figure 1 showcases the snippet of the dataset that was used in our experiments.

B. Data Cleaning, Integration, and Feature Engineering

Before the application of machine learning, it is quite important to make preprocessing. Because without preprocessing, the data will not undergo the required changes and the results might become biased or improper. Firstly, we dealt with missing values in the dataset. We handled missing values of features based on their data types in the dataset. Because if the data type is string (text), then it will have another strategy rather than if the data type of the feature was integer. For instance, we used the mode for categorical features, whereas for numerical features we used median value as a replacement. We set unsubmitted assessment scores to zero. To show that someone did not withdraw, missing unregistration dates were given a value of -1. We used features of id student, code

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.



code_module	code_presentation	id_assessment	assessment_type	date	weight
Identification code of the module, to which the assessment belongs.	Identification code of the presentation, to which the assessment belongs.	Identification number of the assessment.	Type of assessment. Three types of assessments exist: Tutor Marked Assessment (TMA), Computer Marked	Information about the final submission date of the assessment calculated as the number of days since the start of	Weight of the assessment in %. Typically, Exams are treated separately and have the weight 100%;
FFF 25% BBB 20% Other (112) 54%	2014J 28% 2014B 28% Other (92) 45%	1752 40.1k	TMA 51% CMA 37% Other (24) 12%	12 261	0 100
AAA	2013J	1752	TMA	19	10
AAA	2013J	1753	TMA	54	20
AAA	2013J	1754	TMA	117	20
AAA	2013J	1755	TMA	166	20
AAA	2013J	1756	TMA	215	30
AAA	2013J	1757	Exam		100

Fig. 1: Snippet of dataset.

module, and code presentation to combine the tables. As a result, we put together features as total and average assessment scores, total VLE clicks, normalized clicks per credit, and score per assessment. Other engineered features are the time between registration and the start of a module and the length of time a person can participate.

C. Train-Test Partitioning

Division of the dataset into train and test is very important. Because even if the accuracy is high in training, it does not mean that the accuracy will be high in testing part. The 80/20 stratified split was applied to preserve class distribution for classification tasks. Stratification ensures reliable performance estimate by preventing over or under presence of any class. Two main models were used in the work: logistic regression and random forest. At the initial stage, five classical algorithms were considered: LR, RF, Gradient Boosting, AdaBoost, and the support vector machine method. LR and RF were selected for further analysis, as they demonstrated a higher quality of forecasting, better interpretability and stability of the results.

D. Models

This section explains the models that we used in our study with corresponding justification of uses. To train logistic regression, a stratified 5-fold cross-validation was used with the adjustment of hyperparameters C and solver. The evaluation metrics used were macro-accuracy, macro-completeness, macro F1-measure, and accuracy. Confusion matrices were visualized using the `seaborn` library for visual analysis of the results. Logistic regression served as a transparent linear basic model reflecting the general patterns of engagement levels. An accelerated workflow based on `RandomizedSearchCV` was used to train the random forest model. This approach sampled a smaller number of hyperparameter combinations across `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`, `max_features`, and `bootstrap`. In order to reduce the calculation time and increase the generalizability of the results, a stratified triple cross-validation was used. The same metrics were used for estimation as for logistic regression, and confusion matrices were visualized using `Matplotlib`. This approach simplified the configuration of the model and allowed us to take into account non-linear interactions and a complex data structure. The quality of the model was assessed by accuracy, macro-average values of completeness and F1-measure, and the analysis of confusion matrices allowed us to determine the nature of the forecast errors for each class. The combination of these indicators provided a comprehensive assessment of the effectiveness of the model in predicting the level of student engagement.

IV. Results and Discussion

This section showcases obtained results and the discussion of what we found with corresponding analysis of strengths and weaknesses of our obtained results. To make hyperparameter tuning we applied grid search. It was applied to aim to get optimal values for hyperparameters of the applied models. The logistic regression model was configured using a grid search to determine optimal hyperparameter values:

```
C = 1, solver = 'liblinear'
```

The performance indicators of the logistic regression model are shown in table I. The overall accuracy of the model was determined at 0.896, while the macro-average indicators of accuracy, completeness and F1 dimensions were also approximate to this value. The classification report (Table II)indicates balanced results for both classes.

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

TABLE I: Evaluation Metrics (Weighted & Macro) of Logistic Regression

Metric	Score
Accuracy	0.8957
Precision (Macro)	0.8959
Recall (Macro)	0.8972
F1-score (Macro)	0.8956

TABLE II: Classification Report of Logistic Regression

Class	Precision	Recall	F1-score	Support
0	0.93	0.87	0.90	3442
1	0.86	0.92	0.89	3077
Accuracy		0.90		
Macro Avg	0.90	0.90	0.90	6519
Weighted Avg	0.90	0.90	0.90	6519

The results show that logistic regression provides a fairly accurate classification of student engagement, demonstrating a slight improvement in completeness for class 1 and a slight increase in accuracy for class 0. The following optimal hyperparameter values were determined for the random forest model (accelerated version):

```
n_estimators = 160, min_samples_split = 10, min_samples_leaf = 3,
max_features = None, max_depth = 15, bootstrap = True
```

In general, the random forest model surpassed the logistic regression, reaching an accuracy of 0.926. The macro-averaged values of accuracy, completeness, and F1-measure also amounted to about 0.926 (Table III). The classification report presented in IVshows stable performance of the model in both classes, with a slight increase in accuracy for class 1 and completeness for class 0 compared to logistic regression.

TABLE III: Random Forest Performance Metrics

Metric	Score
Accuracy	0.9262
Precision (Macro)	0.9265
Recall (Macro)	0.9278
F1-score (Macro)	0.9262

TABLE IV: Random Forest Classification Report

Class	Precision	Recall	F1-score	Support
0	0.96	0.90	0.93	3442
1	0.90	0.96	0.92	3077
Accuracy		0.93		
Macro Avg	0.93	0.93	0.93	6519
Weighted Avg	0.93	0.93	0.93	6519

The findings demonstrate that Random Forest, employing an ensemble-based approach, markedly improves both accuracy and class performance equilibrium in comparison to Logistic Regression. This study did not examine additional algorithms such as Gradient Boosting, AdaBoost, and Support Vector Machines, as the integration of Logistic Regression and Random Forest demonstrated significant efficacy and sufficient benchmarking for traditional supervised methods. We could focus on making sure our results could be repeated, that our models were easy to understand, and that we were using well-known metrics for a strong comparison

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

with this method. We created confusion matrices for both Logistic Regression (Figure 1) and Random Forest (Figure 2) to see how well the models worked. These matrices show in great detail how well each model groups students based on how interested they are. The confusion matrix shows that Logistic Regression worked about the same for both classes. Of the 3442 students who were either low-engaged or failed, 93 percent were correctly classified, and 7 percent were incorrectly classified as high-engaged or passed. In the high-engagement/pass category, 86% were accurately identified, whereas 14% were erroneously predicted to exhibit low engagement/failure. This pattern shows that Logistic Regression does a good job of finding overall trends, but it does show a slightly higher rate of false negatives in the high-engagement group.

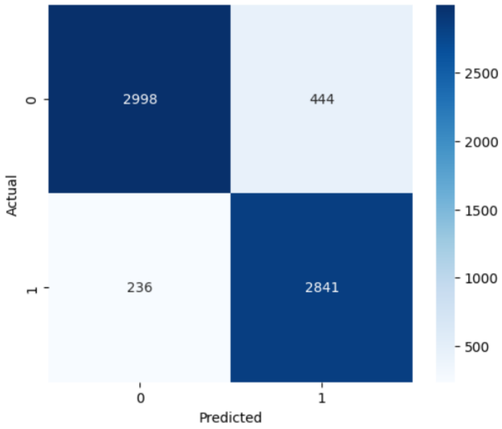


Fig. 2: Confusion matrix for Logistic Regression.

The Random Forest confusion matrix shows that it can now classify things much better. Ninety-six percent of the time, students who did not participate or failed were correctly identified. Ninety percent of the time, students who did participate or passed were correctly identified. Both classes make fewer mistakes than Logistic Regression. This shows that the model can find complicated patterns and interactions that are not straight lines in the feature set. In general, Random Forest works better across classes and makes fewer mistakes, which is what you would expect from its higher F1-score and accuracy. These confusion matrices show how important it is to perform preprocessing and feature engineering. Both models can tell how engaged someone is by making meaningful behavioral features and making engagement signals across modules the same. The more complex, nonlinear interactions in the dataset are what make Random Forest work best.

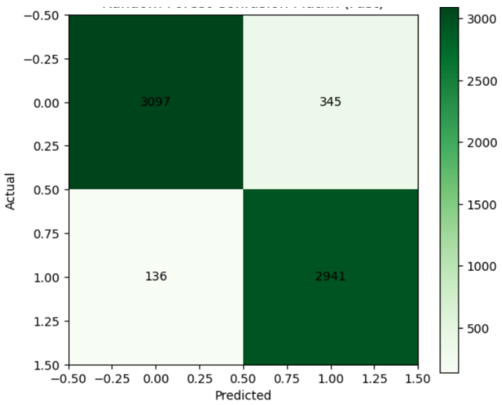


Fig. 3: Confusion matrix for Random Forest.

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

V. Conclusions and Future Work

This study shows that classical machine learning models can effectively predict student engagement in online learning environments when used with a carefully designed preprocessing and feature engineering pipeline. We used the Open University Learning Analytics Dataset (OULAD) in order to get behavioral indicators from clickstream and assessment data, which made it easier to classify engagement in a more complete way. We looked at Logistic Regression and Random Forest, and we found that Random Forest showed better accuracy and performance in all engagement categories. These results show that traditional supervised learning methods can still be useful for predict engagement when they are used with careful preprocessing. They can also be considered as clear, repeatable standards for research in learning analytics. This study shows that the preprocessing of data and the formation of characteristic features have a decisive influence on the quality of the model. Converting the source information from the LMS system into structured capabilities will enhance the interpretation of models and increase the accuracy of forecasting. Such approach is important in the educational process for the early identification of risk groups and the development of effective training strategies. The proposed method can be adapted to various educational data sets and conditions for conducting comparative analysis. In upcoming studies, the use of time modeling, as well as the introduction of cognitive and emotional indicators obtained through questionnaires and computer vision methods, can further improve the quality of the forecast. In addition, a comparison of classical machine learning approaches with modern models of deep learning paves the way for a deeper understanding of the relationship between interpretation capability and computational complexity.

References

[1] Jakub Kuzilek, Martin Hlosta, Zdenek Zdrahal, *Open University Learning Analytics dataset*, Scientific Data, vol.4, p.170171, 2017. DOI: https://doi.org/10.1038/sdata.2017.171

[2] Khurram Jawad, Muhammad Arif Shah, Muhammad Tahir, *Students' Academic Performance and Engagement Prediction in a Virtual Learning Environment Using Random Forest with Data Balancing*, Sustainability, vol.14, no.22, p.14795, 2022. DOI: https://doi.org/10.3390/su142214795

[3] M. Yağcı, *Educational data mining: prediction of students' academic performance using machine learning algorithms*, Smart Learning Environments, vol.9, p.11, 2022. DOI: https://doi.org/10.1186/s40561-022-00192-z

[4] Zhaoyu Shou, Mingquan Xie, Jianwen Mo, Huibing Zhang, *Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach*, Applied Sciences, vol.14, no.6, p.2522, 2024. DOI: https://doi.org/10.3390/app14062522

[5] M. A. A. Dewan, M. Murshed F. Lin, *Engagement detection in online learning: a review*, Smart Learning Environments, vol.6, article 1, 2019. DOI: https://doi.org/10.1186/s40561-018-0080-z

[6] N. A. Johar, S. N. Kew, Z. Tasir E. Koh, *Learning Analytics on Student Engagement to Enhance Students' Learning Performance: A Systematic Review*, Sustainability, vol.15, no.10, article 7849, 2023. DOI: https://doi.org/10.3390/su15107849

[7] M. B. Garcia, C.-L. Goi, K. Shively D. Maher, *Understanding Student Engagement in AI-Powered Online Learning Platforms: A Narrative Review of Key Theories and Models*, in Cases on Enhancing P-16 Student Engagement With Digital Technologies, 2025. DOI: https://doi.org/10.4018/979-8-3693-5633-3.ch001

[8] S. S. Khan, A. Abedi T. Colella, *Inconsistencies in the Definition and Annotation of Student Engagement in Virtual Learning Datasets: A Critical Review*, arXiv preprint, 2022. URL: https://arxiv.org/abs/2208.04548

[9] S. N. Karimah et al., *Automatic engagement estimation in smart education/learning settings: a systematic review of engagement definitions, datasets, and methods*, Smart Learning Environments, vol.9, article31, 2022. DOI: https://doi.org/10.1186/s40561-022-00212-y

[10] N. J. Falkner K. E. Falkner, *Predicting Student Engagement in the Online Learning Environment*, International Journal of Web-Based Learning and Teaching Technologies, vol.16, pp.1–17, 2021. DOI: https://doi.org/10.4018/IJWLTT.287095 (sciencedirect.com)

[11] M. A. Al Mamun G. Lawrie, *Student-content interactions: Exploring behavioural engagement with self-regulated inquiry-based online learning modules*, Smart Learning Environments, vol.10, article1, 2023. DOI: https://doi.org/10.1186/s40561-022-00221-x (slejournal.springeropen.com)

[12] B. Flanagan, R. Majumdar H. Ogata, *Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis*, International Journal of Educational Technology in Higher Education, vol.19, article41, 2022. DOI: https://doi.org/10.1186/s41239-022-00348-4 (educationaltechnologyjournal.springeropen.com)

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.

[13] User667, *OULAD Personalized Learning Path*, Kaggle Notebook and Dataset, 2025. URL: https://www.kaggle.com/code/user667/oulad-personalized-learning-path

[14] P.-C. Sun, R. J. Tsai, G. Finger, Y.-Y. Chen D. Yeh, *What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction*, Computers Education, vol.50, pp.1183–1202, 2008. DOI: https://doi.org/10.1016/j.compedu.2006.11.002

[15] C. R. Henrie, L. R. Halverson C. R. Graham, *Measuring student engagement in technology-mediated learning: A review*, Computers Education, vol.90, pp.36–53, 2015. DOI: https://doi.org/10.1016/j.compedu.2015.09.005

[16] M. Bond, S. Bedenlier, V. I. Marín M. Händel, *Student engagement in blended learning environments: A review of the literature*, International Journal of Educational Technology in Higher Education, vol.17, article1, 2020. DOI: https://doi.org/10.1186/s41239-020-00214-5

[17] J. A. Fredricks, P. C. Blumenfeld A. H. Paris, *School engagement: Potential of the concept, state of the evidence*, Review of Educational Research, vol.74, pp.59–109, 2004. DOI: https://doi.org/10.3102/00346543074001059

[18] R. M. Carini, G. D. Kuh S. P. Klein, *Student engagement and student learning: Testing the linkages*, Research in Higher Education, vol.47, pp.1–32, 2006. DOI: https://doi.org/10.1007/s11162-005-8150-9

[19] Qi, Y.; Zhuang, L.; Chen, H.; Han, X.; Liang, A. *Evaluation of Students' Learning Engagement in Online Classes Based on Multimodal Vision Perspective*, Electronics, vol.13, article149, 2024. DOI: https://doi.org/10.3390/electronics13010149 (mdpi.com)

[20] Tang, X.; Gong, Y.; Xiao, Y.; Xiong, J.; Bao, L. *Facial Expression Recognition for Probing Students' Emotional Engagement in Science Learning*, Journal of Science Education and Technology, vol.34, pp.13–30, 2025. DOI: https://doi.org/10.1007/s10956-024-10143-7 (link.springer.com)

[21] Kuzilek, J., Hlosta, M. & Zdrahal, Z. Open University Learning Analytics dataset. *Scientific Data* **4**, 170171 (2017). https://doi.org/10.1038/sdata.2017.171

Received: December 28, 2025. Reviewed: December 29, 2025. Accepted: December 30, 2025.