Improving the Effectiveness of E-learning Videos by leveraging Eye-gaze Data

Rakhi Saxena
Department of Computer Science, Deshbandhu College, University of Delhi, India
rsaxena@db.du.ac.in

Sunita Narang
Department of Computer Science, Acharya Narendra Dev College, University of Delhi, India
sunitanarang@andc.du.ac.in

Harita Ahuja
Department of Computer Science, Acharya Narendra Dev College, University of Delhi, India
haritaahuja@andc.du.ac.in (corresponding author)

Received: 5 September 2023 | Revised: 26 September 2023 | Accepted: 4 October 2023

 DOI: https://doi.org/10.48084/etasr.6368

ABSTRACT

Recent advances in technology strengthen remote and lifelong learning by integrating e-videos into teaching-learning pedagogy. Therefore, educational content developers are tasked with creating engaging and qualitative e-content. The shift in paradigm from offline to online teaching brings forth several issues regarding the quality of online learning materials and the missing dynamic interaction between instructors and learners. Leveraging contemporary artificial intelligence techniques to provide insights into methods for developing quality e-content is the need of the hour. This study showed that the pattern and duration of the eye gaze of the learner on the text, image, or instructor in the video reveal valuable insights, not only regarding the comprehension of the learner but also giving suggestions to improve video lectures. The results show that learners perform better when they spend more time looking at the instructor compared to the image and text on a frame. Therefore, just like classroom teaching, the presence of the instructor in the video is vital, as looking directly at the instructor while they are delivering the lecture encourages comprehension. Furthermore, by applying classification techniques to learner eye gaze data, it was possible to predict with 97% confidence whether the learner would answer the post-quiz correctly or not.

Keywords-e-learning; eye gaze data; prediction; classification; machine learning

I. INTRODUCTION

Education policies around the world place special emphasis on online education to improve teaching and evaluation pedagogies. They envision that technology will play a pivotal role in the improvement of the teaching-learning process, boosting educational access, and integrating e-content into teaching-learning practices [1]. The occurrence of the global pandemic and other crises has demonstrated the necessity for online models to provide education whenever in-person models of education are disrupted [1-2]. E-learning not only aids in information retention but can be adaptive by allowing learners to go through the content using a personalized pedagogical strategy [3]. The transition from offline classroom learning to an online mode of education needs to resolve issues concerning the quality of online learning material, missing dynamic communication between speaker and learners, on-spot learners' query resolution, etc. Thus, the importance of carefully crafted quality e-content that engages the learner and improves comprehension cannot be emphasized enough. The missing interaction between teacher and learner in online video is the main hurdle in the adoption of online learning. Therefore, online education requires engaging and interactive videos to capture the learner's attention [4].

A. Background

Recently, many studies attempted to automatically calibrate the attention and comprehension of the learners by tracking their eye movements while watching online videos [5-6]. Eye tracking is a noninvasive technique that provides vital insights for estimating learner attentiveness [7]. In [8], it was shown that the use of eye-tracking methods has increased substantially since 2001, and that it seems to be a revolutionary method for academics to connect teaching-learning outcomes with cognitive processes [8]. Eye-tracking technology has witnessed a revolutionary development in the last decade due to advances in artificial intelligence, portable electronics, and head-mounted eye-tracker devices [9]. Therefore, researchers have access to equipment to capture eye gaze data efficiently.

B. Motivation

Contemporary MOOCs employ various mechanisms, such as keystrokes, intermediate response time, and prompt responses, to capture the learners' attention, but these are ineffective in accessing the learner's comprehension of the topic. Additionally, though the impact of different ways of including instructors in e-learning videos has been explored in the context of enhancing individuals' learning capabilities [10], the observation that eye movements are synchronized across video sections has not been widely explored in the context of education [4]. Researchers have provided empirical evidence that the presence of an instructor in online videos plays a key role in improving learner cognition, as well as in attracting and retaining students [11-14]. Furthermore, instructor expression, enthusiasm, and interaction can significantly increase the quality of online synchronous learning, leading to knowledge gain by students and improved teacher satisfaction [15-17]. Although all of these studies discussed the positive impact of an instructor's presence on the perceptions of the learners, none of them has leveraged eye gaze data to predict learners' knowledge acquired through watching online videos. Motivated by this, this study aimed to predict learners' takeaways after watching e-content and discuss the role of the instructor's presence on the learner's understanding captured through a post-quiz.

C. Objectives

This study aimed to find the correlation between the duration of learners' eye gaze on specific sections of selected e-learning videos and their comprehension of the subject. The frames of the videos were categorized into sections of text, images, and the instructor. It was hypothesized that the amount of time spent gazing at different segments would reveal the learner's comprehension of the content. The eye gaze data collected were used to predict the comprehension of the learner and provide tips to improve the design structure of the video. Intentional learning and incidental learning are known to affect learning motivation [6]. Keeping this in mind, in the experiments conducted, the learners were told in advance about a quiz that would be presented after watching the video. As the goal was to predict how well learners would perform in the post-quiz related to an instructional video given their eye gaze data, intentional learning was used where stakeholders were a priori informed about the post-quiz.

II. METHODS

This section details the method used for predicting the desired learning outcomes based on the eye gaze data collected. Python 3.6.9 was used for coding purposes and executed on an Intel Core i7-6700 CPU @3.40GHz with 16GB RAM.

A. Experimental Setup

Study participants were shown short instructional videos and asked to attempt a quiz afterwards in order to quantify their comprehension. The OpenCV and Dblib Python libraries were used to track the eye gaze in real-time via a webcam. The conversion of the video to its corresponding frames was performed using the OpenCV library. A fixation of eye gaze was used and the spatial coordinates (*x_pos*, *y_pos*) of the area of interest were noted along with fixation_duration, indicating the time for which the *x_pos* and *y_pos* were maintained by the eye gaze.

B. Data Description

Four short videos were selected. In each of these videos, the instructor was present on the screen, and the rest of the screen consisted of text and/or images. Table I shows the duration of videos screened and the number of participants. Eye gaze data were collected from the cohort of learners while they watched these videos.

TABLE I. COHORT SCREENING DATA AND DURATION TIME

Video	Duration	Number of participants
Video1	4:59	50
Video2	4:39	65
Video3	4:56	63
Video4	4:40	46

C. Data Cleaning and Preprocessing

The data were cleaned by eliminating erroneous data points due to eye fixation on regions outside the video frame. One possible reason for outlier data points can be attributed to participants moving their heads while watching the video, even though they were informed to keep them stable. The videos were split into frames and manually designated regions of each frame as text, image, and instructor by drawing bounding boxes around the respective region. The collected eye fixation point data were segregated into categories (*text*, *image*, or *instructor*) depending on the specific bounded box in which the *x_pos* and *y_pos* coordinates lie. In case eye gaze data were from neither region, it was marked as *others* (outliers). In some frames, the eye gaze data points belonged to overlapping regions, and these points were classified as parts of both categories.

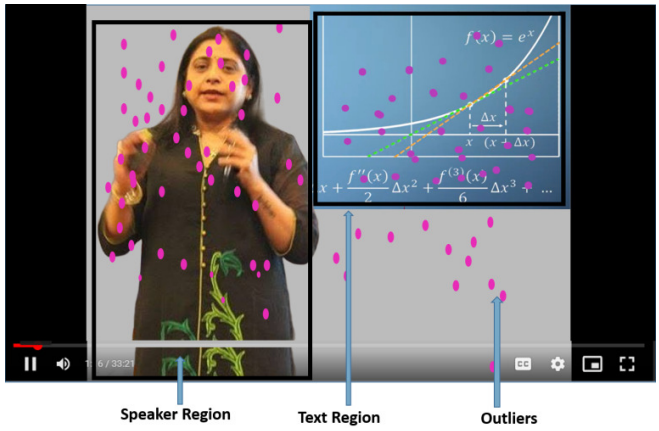

Fig. 1. Bounding boxes categorizing *text*, *image*, *instructor*, *overlapping*, and *others* regions in the frame and data points mapped to each category.

Figure 1 shows a sample frame with bounded boxes and the distribution of eye gaze coordinates in those regions. Each segment is marked with its name in the figure for clarity. This frame shows only a section of the data for demonstration purposes. Figure 2 shows a plot of the frequency distribution of the eye gaze data corresponding to the four categories for one frame.

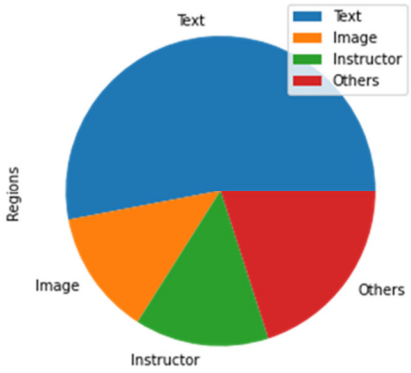

Fig. 2. Percentage-wise frequency distribution of features per frame.

D. Feature Engineering

To build a model to predict the post-quiz results, features for questions were derived using fixation duration in respective eye gaze categories. For each question asked in the quiz, the frame having the information required to answer the question was identified. The feature score for the question was computed as the fraction of the duration for which the student looked at the particular category. As a final step, the scores were normalized so that the sum for each feature was equal to 1 for each question. Normalization of the scores was required for each question due to the different number of total frames processed for each question. Figure 3 shows the data distribution and the five-number statistics summary for the normalized features.

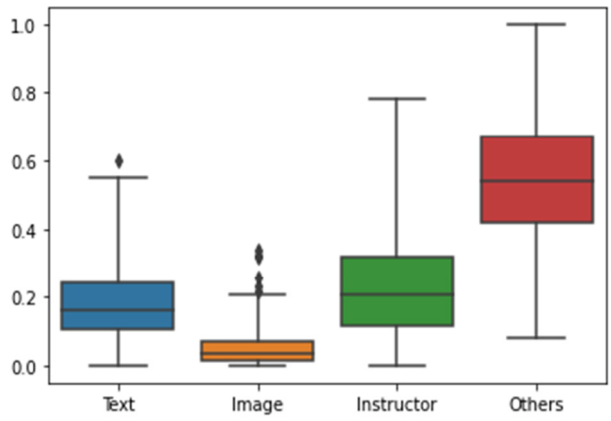

Fig. 3. Visualization of five-number statistics of features.

The five-number statistics summary is a set of descriptive statistics that includes the values of the sample minimum, maximum, median, first quartile, and third quartile. Algorithm 1 shows the steps to calculate the scores for each question.

```
Algorithm 1: Calculate feature scores for questions
Result: Feature Scores
for every question do
  Set text=image=instructor=others = 0
  find frame_start and frame_end
  if fixation_point ∈ text_bounding_box then
    text + = current_fixation_duration
  if fixation_point ∈ image_bounding_box then
    image + = current_fixation_duration
  if fixation_point ∈ instrctr_bounding_box then
    instructor += current_fixation_duration
  else
    others += current_fixation_duration;
  end
return text, image, instructor, others
```

III. SUPERVISED PREDICTION MODELS

The following machine learning classification models [18] were used with default hyperparameters to predict the learner's response using their feature scores computed through the eye gaze data. Note that in experimentation, an 80:20 ratio of the data was used for training and testing, respectively, for building the classifiers and predicting the results. The reported results were averaged over 20 runs.

- Gradient Boosting Classification*:* This machine learning technique is generally used for regression and classification problems and produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.
- Support Vector Machine (SVM): This supervised learning model analyzes the data used for classification and regression analysis. Given a set of labeled training examples, an SVM builds a model that assigns new examples to one category or another using a decision boundary for the classes.
- Neural Network (NN): This model is inspired by the structure and workings of the human brain. The network consists of multiple node layers that have an input layer, multiple hidden layers, and a single output layer. It learns to perform tasks by considering examples, generally without being programmed with any task-specific rules.

IV. RESULTS

Feature scores for all questions were tallied with the answers submitted by the learners in the post-quiz (binary variable *answer*) using correlation statistics to determine the extent of relationships between the two variables. The Pearson standard correlation coefficient was used to determine the degree and direction of the relation between the features and the quiz scores. Pearson correlation is calculated using least-squares and is defined as the ratio between the covariance of the variables and the product of their standard deviations as:

$$\rho = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

where $cov(x, y)$ is the covariance, σ_x is the standard deviation of x and σ_y is the standard deviation of y. Note that, ρ = 1 represents a perfect positive relationship, ρ = -1 denotes a perfect negative relationship, and ρ = 0 indicates the absence of a relationship between the variables.

A. Pearson Coefficient of Correlation to Identify Influential Features

Figure 4 shows the Pearson's correlation coefficient for the complete data through a heat map, indicating that there is a

negative correlation between the post-exam score field (*answer*) and the *text*, as well as the *image* field. This negative correlation makes the formulated assumption void. However, a positive correlation between the *answer* and the *instructor* field reflects that learners' responses are influenced by the instructor feature. To further validate this hypothesis, the p-values were calculated, as shown in Table II, while testing the significance of these correlation coefficients. It was found that there was a significant relationship between the instructor feature and the answer, and the result of the experiment was statistically significant (97%). The results corresponding to the rest of the features were not statistically significant.

Fig. 4. Heat map showing the correlations between features and answer.

TABLE II. P-VALUES FOR SIGNIFICANCE TESTING

p-value	Text	Image	Instructor	Others
Answer	0.20	0.358	0.034	0.96

B. Evaluation Metrics for Classification Models

Multiple classification models were run with default hyperparameters, and the following evaluation metrics were computed: Accuracy, Precision, Mean Squared Error (MSE), and Mean Absolute Error (MAE) [18]. The Python Scikit-learn (Sklearn) library was used to run various multiclass classifiers [19]. Six classifiers from the library were used: Gradient Boosting Classifier (GBRegression) [20], Epsilon Support Vector Regression (EVRegression) [21], Kernel Ridge Regression (KRRegression) [22], Linear Support Vector Regression (LinearSVR) [21], and two-layered NN (2NN) [23]. Classifiers were run using the Sigmoid and ReLu activation functions. Linear Support Vector Regression was also performed after applying Principal Component Analysis (PCA) on the feature space (PCA+SVR) [21].

Figure 5 shows the comparative bar plot for the all metrics for the six classifiers, indicating that GBRegression, PCA+SVR, and NN delivered the best performance with the least errors. Table III shows different SVM kernels used to see the variations in errors. It was found that PCA+SVR with kernel "poly" resulted in the least errors among the three kernels used. However, the lowest error values of GBRegression confirm its efficacy compared to SVMs for the considered dataset.

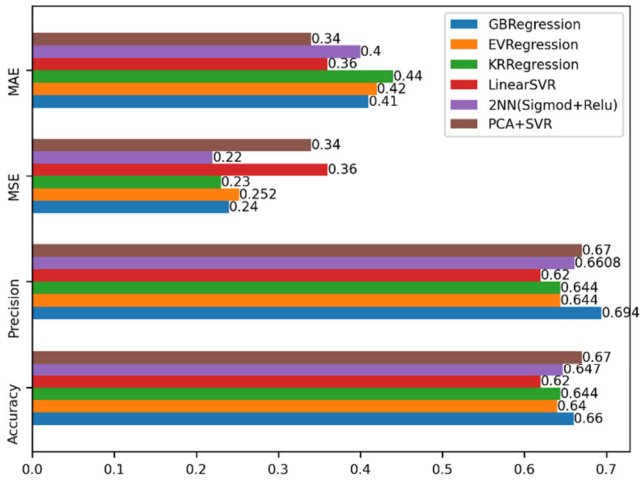

Fig. 5. Comparative evaluation of multiple classifiers using different evaluation metrics (shown as x-axis).

TABLE III. ERRORS FOR GRADIENT BOOSTING CLASSIFIERS AND SVM WITH DIFFERENT KERNELS

Classifier	MSE	MAE
Gradient Boosting	0.23	0.41
PCA+SVR (Kernel = "Linear")	0.241	0.407
PCA+SVR (Kernel = "Poly")	0.239	0.4
PCA+SVR (Kernel = "RBF")	0.238	0.412

C. Role of Optimizers in NN

In addition, 2NNs were run with different epochs and optimizers to see the variations in the accuracy of the built model. Note that an optimizer is a function that updates the attributes, such as weights and learning rate, of the NN, reduces the overall loss, and improves accuracy. Figure 6 shows the achieved accuracy, indicating that a 2NN with an SGD optimizer function resulted in the highest accuracy. Since it is an ensemble-based method, it builds the model in a stage-wise fashion, and, hence, it generalizes better compared to other models, thus leading to a higher test-set accuracy.

D. Role of Features on Learner's Comprehension

Table II and Figure 4 show that *instructor* is the most significant attribute because it is highly correlated with *answer* given by the individuals in the post-tests. Hence, the question "Why not use only this feature for model building instead of using all four features as considered in former experiments?" arises. To show the importance of all features in model building, the following cases were considered as three different feature sets, and the best-supervised model (GBRegression) was used to compare their performances using precision and accuracy metrics:

1. Use the *instructor* as the only feature.
2. Remove *instructor* from the feature set, i.e., use three features *text*, *image*, and *others* in model building to predict the learner's understanding through e-content.
3. Use all features.

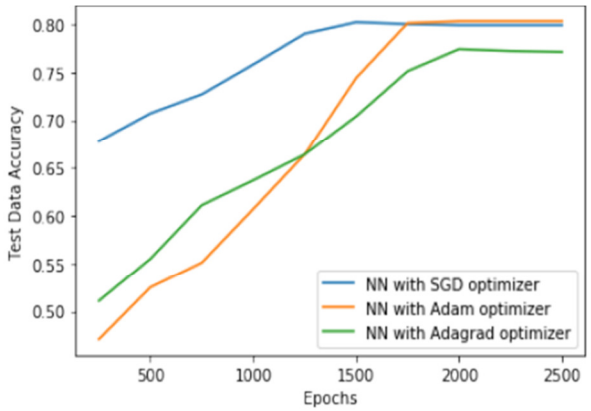

Fig. 6. Accuracy of a 2NN with different optimizers and epochs.

Figure 7 shows the precision and accuracy metrics of the built models. As the values of the metrics were higher in case 3 where all features are used, all features are important for the learner's comprehension. Similar observations were made with the rest of the features individually. However, the increase in precision and accuracy was greater after including *instructor* compared to the rest of the features. Hence, *instructor* plays an important role in learning comprehension.

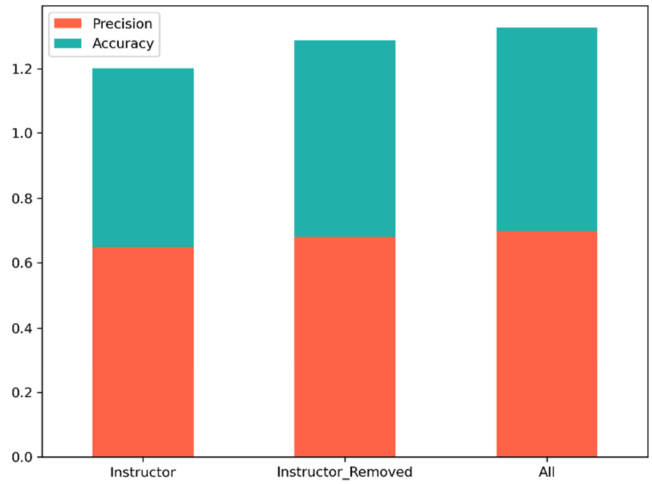

Fig. 7. Performance metrics precision and accuracy of GBRegression model for different cases.

V. DISCUSSION

The analysis of the experiments concludes that learners obtained improved post-exam scores in the instances when they spent more time looking at the speaker compared to the visual and textual data for a particular question. There is a relatively weaker negative correlation between the *image* feature and the post-exam score (*answer*). Similarly, a relatively stronger negative correlation was also observed between the *text* score and *answer*. It was also noted that there was a much stronger positive correlation between the *instructor* score and the student's post-exam score (*answer*). Thus, looking directly at the teacher, while he/she is delivering the lecture content, results in better learning than by looking at the text/image contents only. This conclusion was also confirmed in experiments conducted using different combinations of features.

These results also show that whether the student would get a particular question right or wrong can be predicted with nearly 70% accuracy from the engineered features. This confirms the potential of the features used in model training. After trying a variety of classification models, a 2NN model with an SGD optimizer delivered the highest classification accuracy on the test dataset.

VI. CONCLUSION

This study aimed to find the correlation between the duration of learners' eye gaze on specific sections of selected e-learning videos and their comprehension. The frames of the e-learning videos were categorized into sections of text, images, the instructor, and others. Three supervised prediction models were used along with six classification optimizers. The results showed that the learners obtained improved post-quiz scores when they spent more time looking at the speaker compared to the visual and textual data for a particular question. When comparing the evaluation metrics of the models used, a two-layered NN with an SGD provided the highest classification accuracy on this task.

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of Mayank Saxena, Delhi Technological University, India for the collection of the eye gaze data.

REFERENCES

[1] R. Govinda, "NEP 2020: A Critical Examination," *Social Change*, vol. 50, no. 4, pp. 603–607, Dec. 2020, https://doi.org/10.1177/0049085720958804.

[2] H. Chatti and S. Hadoussa, "Factors Affecting the Adoption of E-Learning Technology by Students during the COVID-19 Quarantine Period: The Application of the UTAUT Model," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6993–7000, Apr. 2021, https://doi.org/10.48084/etasr.3985.

[3] H. E. Fazazi, M. Elgarej, M. Qbadou, and K. Mansouri, "Design of an Adaptive e-Learning System based on Multi-Agent Approach and Reinforcement Learning," *Engineering, Technology & Applied Science Research*, vol. 11, no. 1, pp. 6637–6644, Feb. 2021, https://doi.org/10.48084/etasr.3905.

[4] J. Madsen, S. U. Júlio, P. J. Gucik, R. Steinberg, and L. C. Parra, "Synchronized eye movements predict test scores in online video education," *Proceedings of the National Academy of Sciences*, vol. 118, no. 5, Feb. 2021, Art. no. e2016980118, https://doi.org/10.1073/pnas.2016980118.

[5] A. R. Khan, S. Khosravi, S. Hussain, R. Ghannam, A. Zoha, and M. A. Imran, "EXECUTE: Exploring Eye Tracking to Support E-learning," in *2022 IEEE Global Engineering Education Conference (EDUCON)*, Tunis, Tunisia, Mar. 2022, pp. 670–676, https://doi.org/10.1109/EDUCON52537.2022.9766506.

[6] Y. Y. Ng and A. Przybyłek, "Instructor Presence in Video Lectures: Preliminary Findings From an Online Experiment," *IEEE Access*, vol. 9, pp. 36485–36499, 2021, https://doi.org/10.1109/ACCESS.2021.3058735.

[7] J. Wang, "Leveraging Eye Tracking Technology to Improve Teacher Education," *Journal of Technology and Teacher Education*, vol. 30, no. 2, pp. 253–264, Aug. 2022.

[8] M. L. Lai *et al.*, "A review of using eye-tracking technology in exploring learning from 2000 to 2012," *Educational Research Review*, vol. 10, pp. 90–115, Dec. 2013, https://doi.org/10.1016/j.edurev.2013.10.001.

[9] M. Q. Khan and S. Lee, "Gaze and Eye Tracking: Techniques and Applications in ADAS," *Sensors*, vol. 19, no. 24, Jan. 2019, Art. no. 5540, https://doi.org/10.3390/s19245540.

[10] M. Abdous and C. J. Yen, "A predictive study of learner satisfaction and outcomes in face-to-face, satellite broadcast, and live video-streaming learning environments," *The Internet and Higher Education*, vol. 13, no. 4, pp. 248–257, Dec. 2010, https://doi.org/10.1016/j.iheduc.2010.04.005.

[11] C. Ilioudi, M. N. Giannakos, and K. Chorianopoulos, "Investigating differences among the commonly used video lecture styles," in *Proceedings of the Workshop on Analytics on Video-based Learning, WAVe*, Leuven, Belgium, 2013, vol. 13, pp. 21–27, https://doi.org/10.13140/2.1.3524.9284.

[12] Y. Zhang, K. Xu, Z. Pi, and J. Yang, "Instructor's position affects learning from video lectures in Chinese context: an eye-tracking study," *Behaviour & Information Technology*, vol. 41, no. 9, pp. 1988–1997, Jul. 2022, https://doi.org/10.1080/0144929X.2021.1910731.

[13] Y. Tao, G. Zhang, D. Zhang, F. Wang, Y. Zhou, and T. Xu, "Exploring Persona Characteristics in Learning: A Review Study of Pedagogical Agents," *Procedia Computer Science*, vol. 201, pp. 87–94, Jan. 2022, https://doi.org/10.1016/j.procs.2022.03.014.

[14] A. T. Stull, L. Fiorella, and R. E. Mayer, "An eye-tracking analysis of instructor presence in video lectures," *Computers in Human Behavior*, vol. 88, pp. 263–272, Nov. 2018, https://doi.org/10.1016/j.chb.2018.07.019.

[15] Z. Pi, K. Xu, C. Liu, and J. Yang, "Instructor presence in video lectures: Eye gaze matters, but not body orientation," *Computers & Education*, vol. 144, Art. no. 103713, Jan. 2020, https://doi.org/10.1016/j.compedu.2019.103713.

[16] Z. Katai and D. Iclanzan, "Impact of instructor on-slide presence in synchronous e-learning," *Education and Information Technologies*, vol. 28, no. 3, pp. 3089–3115, Mar. 2023, https://doi.org/10.1007/s10639-022-11306-y.

[17] Q. Huangfu, H. Li, S. Tang, J. Wang, Q. Liu, and G. Chen, "How teacher enthusiasm affects students' learning of chemistry declarative knowledge in video lectures," *Chemistry Education Research and Practice*, vol. 23, no. 4, pp. 898–912, 2022, https://doi.org/10.1039/D2RP00095D.

[18] M. Alghobiri, "A Comparative Analysis of Classification Algorithms on Diverse Datasets," *Engineering, Technology & Applied Science Research*, vol. 8, no. 2, pp. 2790–2795, Apr. 2018, https://doi.org/10.48084/etasr.1952.

[19] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[21] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. Barlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 2019.

[22] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

[23] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Chia La-guna Resort, Italy, Mar. 2010, pp. 249–256.