

Prediction of Students' Behavior in Online Learning through Body Language Using Machine Learning

Rawad Abdulghafor 

Faculty of Computer Studies (FCS), Arab Open University–Oman, Muscat, Oman
Email: rawad.a@aou.edu.om

Abstract—In online learning settings, the instructor is usually not able to monitor the non-verbal behavior of students in terms of body language and facial orientation, which could be used to tell whether the student is paying attention or not. This paper examines how machine learning can be used to categorize visual cues of body language that can be used to predict attentiveness in students during online learning. The face-pose image dataset was publicly available and obtained through Kaggle and consisted of around 2000 images and served as a proxy dataset of attention-related visual behavior. Image classification was performed using four machine learning models: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Convolutional Neural Network (CNN). The images all underwent preprocessing of resizing to a 48×48 -pixel grayscale to have a consistent input representation. The experimental results indicate that the SVM model achieved the highest classification accuracy of 95%, followed by DT (89%), RF (85%), and CNN (72%). The results indicate that machine learning models have the potential to be effective in classifying body-orientation cues based on image data and can be used to build future systems that will analyze student engagement in online learning environments.

Keywords—online learning, body language, student behavior prediction, student engagement detection, machine learning, image classification

I. INTRODUCTION

Students are now characterized by short attention spans, and therefore, it is of importance to monitor their activity when undertaking learning activities. Although online learning offers flexibility and accessibility, it can also present adaptation challenges for many students, particularly in the early stages of using digital learning environments. Even though modern learners are generally familiar with technology, using online platforms for academic purposes may still create motivational and organizational difficulties. With appropriate guidance and support from instructors, however, students can gradually adapt to these virtual learning environments. The transition from traditional face-to-face classrooms to

computer-based online learning significantly changes the way students interact with instructors and course materials.

Considering body language is regarded as the most important aspect in human communication, it would be interesting to develop a system that can evaluate our posture and provide feedback over time. Students' aversion to change prevents them from adapting to the online learning environment, whilst getting used to computer-based teaching techniques takes time. Passive listening and taking notes are necessary in a traditional classroom, but online discussions or the creation of an online class demand active participation. Students who have a "traditional" mindset have a difficult time transitioning; nonetheless, they must approach the new learning environment with an open mind and heart. Automatic identification of affect/emotions by monitoring facial expression using machine learning computational models has a long history [1]. If students are constantly not listening in class and not paying attention to lectures, the instructors can apply machine learning methods to find the patterns of disengagement and take necessary measures. Instead of challenging students, the instructors can start a supportive conversation to learn about the possible challenges to their participation. These problems can involve personal issues or inability to comprehend course assignments and course materials. With the early detection of such problems, the instructor is able to make better changes to the course design and eliminate obstacles that can impede student learning and engagement. Engagement detection systems that rely on machine learning have already been demonstrated to assist instructors in the analysis of behavioral and visual indicators of students during online learning lessons [2].

Predicting student engagement and attentiveness is an important goal in modern education because it can help instructors identify learners who may require additional support and can improve the design of online teaching strategies. However, building such systems requires datasets and computational methods that can reliably capture visually observable behavioral cues. Therefore, the focus of this study is not to claim full behavioral

understanding in authentic online classrooms, but rather to investigate whether body-language-related visual cues, particularly head pose and face orientation, can be automatically classified using machine learning methods.

In this work, we evaluate whether images showing different viewing directions can serve as a proxy representation for attention-related body-language patterns that may be relevant to online learning analysis. Based on this objective, the research question is defined as follows: Which machine learning model among Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Convolutional Neural Network (CNN) performs best in classifying body-orientation cues from image data that may be useful for future online student behavior analysis? We also acknowledge that the employed dataset is not collected from real online classroom sessions, and therefore the findings should be interpreted as an initial technical step toward future online learning behavior prediction systems.

II. LITERATURE REVIEW

Body language is a significant non-verbal communication and encompasses facial expression, eye movement, body gestures, posture, and other physical actions that are used to express feelings and intentions, without speaking [3]. In education, non-verbal communication may give significant information on the attentiveness, engagement, and emotional conditions of students, especially in online learning where physical interaction is restricted. Abbas and Khan [4] also stressed the fact that non-verbal communication is closely related to affectivity and may present meaningful emotional and behavioral cues in the course of communication. In the same manner, Sharma and Vyas [5] emphasized the role of interpretation of non-verbal communication in online classes to enhance communication and interaction with students.

Recent studies have covered the application of computer vision and machine learning technologies to study the behavioral patterns of students in online learning settings. Zhang *et al.* [6] were able to suggest a way of identifying the level of involvement of students in the online learning environments through tracking facial expression movement and mouse interaction behavior. Their findings proved that it is possible to determine student engagement automatically in an online learning session. Likewise, Yan and Au [7] also examined the behavior analysis of online learning based on machine learning methods, and they revealed that data-driven models could offer information about the learning patterns and interactions of students in online learning settings.

The study of facial expressions has emerged as a significant research area when it comes to the perception of human feelings and behavioral expressions. Face detection, feature extraction and expression classification were the three basic stages of the facial expression recognition system found by Fasel and Luetttin [8]. The implementation of deep learning methods and especially the CNNs has greatly enhanced the quality of facial emotion recognition systems [9]. Besides this, classical

machine learning solutions together with feature descriptors have performed well. As an example, Revina and Emmanuel [10] showed that local descriptors and Support Vector Machines were effective in facial expression recognition.

The use of machine learning methods to predict student performance and behavioral patterns in school settings has also gained wide usage. Acharya and Sinha [11] explored how soon student performance could be predicted based on a number of machine learning models such as Artificial Neural Networks (ANN), Naive Bayes and SVM. Their results indicated that machine learning models can be used successfully to analyze student data to predict learning outcomes.

Recent works have paid more attention to the identification of student attention and engagement in the online learning environment by visual data. Ngo *et al.* [12] explored how students can be studied using facial expression recognition to determine emotional regulation in the context of online learning in a synchronous collaborative learning setting. Elbawab and Henriques [2] suggested a machine-learning architecture of student attentiveness detection based on webcam images by comparing emotional and non-emotional facial features, which attained promising classification results. On the same note, Pabba and Kumar [13] came up with a multi-cue system that combines facial expressions and head movement, and gaze direction to track classroom activity in smart classroom settings.

Engagement detection has also demonstrated high potential of deep learning methods. Lasri *et al.* [14] developed a deep learning algorithm, which is founded on the Visual Geometry Group 16-layer network (VGG16) architecture, to identify the level of engagement among deaf and hard-of-hearing students. Their model has high classification accuracy with transfer learning methods. Also, systematic reviews like that by Pinto *et al.* [15] emphasized the applicability of convolutional neural networks in tasks of facial expression recognition and emotion detection.

Other recent publications have also addressed the topic of sophisticated emotion recognition models and engagement analysis frameworks. Alhussan *et al.* [16] suggested an optimized Support Vector Machine model in facial expression recognition, and the model has a high classification accuracy. Bisogni *et al.* [17] compared the deep learning techniques with handcrafted feature-based techniques of emotion recognition and revealed that both techniques have their strengths based on the image quality and the environmental factors.

Multimodal systems have been suggested in the context of intelligent educational environments to enhance the analysis of engagement. Irfan *et al.* [18] introduced that the EngageSense system combines gaze direction, facial expression, and body posture in order to determine engagement during virtual classrooms with a high classification accuracy. On the same note, Vani and Jayashree [19] introduced a multimodal emotion recognition system where video, audio and textual

information are integrated to identify emotions through e-learning systems.

Lastly, Mutawa and Sruthi [20] showed that machine learning models could be employed to determine student satisfaction and emotional reactions in the online learning space through survey and interaction data using ANN and RF models.

III. MATERIALS AND METHODS

This paper suggests a machine learning model to categorize body-orientation signals that could be used to determine the attentiveness of students during online learning. The process of the methodology has multiple core steps, such as the selection of the dataset, pre-processing of the images, label preparation, training of the model, and evaluation of the performance. The input data of the experiments was a publicly available face-pose image dataset. The SVM, DT, RF, and CNN classification algorithms were implemented and tested to find out which model gives the best classification performance. The typical evaluation measures that were employed in assessing the effectiveness of the models were accuracy, precision, recall, F1-Score, confusion matrices, and cross-validation. The overall sequence of these methodological stages is illustrated in Fig. 1.

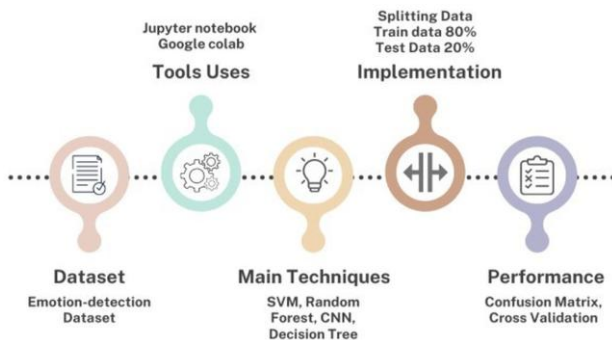


Fig. 1. Overall methodology of the proposed machine learning-based behavior classification system.

This study had the following workflow: dataset selection, image preprocessing, label preparation, model training, hyperparameter tuning, and performance evaluation. The dataset was preprocessed and split into training and testing parts in an 80:20 proportion. Model development was done using the training subset, whereas the evaluation was done using the testing subset, which was not seen before. The training set was hyperparameter tuned and model selected using 5-fold cross-validation during the training stage. Once the optimal settings were chosen in each classifier, the final model was retrained using the entire training set and tested on the held-out test set. Accuracy, precision, recall, F1-Score and confusion matrices were used to compare the performance of the models.

In a bid to enhance reliability and reproducibility of the experimental results, statistical validation procedures were included in the model development. In particular, the training subset was cross-validated 5 times. There were four training data parts in every fold and one part as

validation data and this was repeated five times to ensure that all the folds acted as validation data one time. Once the optimal hyperparameter settings had been identified, each classifier was rerun with the full training set and tested with the independent test set.

To further assess the stability of the obtained results, repeated experimental runs were considered during the evaluation process. Although the reported results are based on a fixed train-test split with cross-validation during training, slight variations in model performance may occur due to data partitioning and model initialization. Future work will include multiple independent runs and statistical reporting of mean and standard deviation values to provide a more comprehensive evaluation of model robustness.

A. Dataset

One of the major challenges in visual behavior analysis research is obtaining suitable datasets that are sufficiently labeled and usable for classification experiments. In this study, the Face Pose Dataset available on Kaggle (<https://www.kaggle.com/datasets/nducloi107/face-pose-dataset>) was used for the experiments, and contains approximately 2000 facial and upper-body images. The images represent different human pose/view categories, such as looking left, right, upward, downward, and straight toward the camera. In the context of this work, these visual categories were used as proxy indicators of body-language-related attention cues. It should be emphasized that the dataset contains generic human head-pose images rather than recordings of students in real online learning environments. Therefore, in this study these pose categories are treated as proxy indicators of attention-related body orientation cues rather than verified measurements of actual student behavior.

It is important to note that the dataset was not collected from real online learning sessions. Therefore, it does not directly represent actual student behavior in virtual classrooms. Instead, it provides a controlled image set that allows us to investigate the feasibility of automatically classifying posture- and orientation-related visual patterns that could be relevant to online engagement analysis in future systems.

To label the images, the images were classified into the corresponding pose / view category of the original dataset. These class labels were utilized as target classes in the supervised learning. The data were split into an 80:20 proportion (training and testing), where 80% of the images were used to develop the model and 20% were held back to evaluate the final results. In the training subfield, hyperparameter tuning and model selection were done using 5-fold cross-validation. The names of the labels in the given work are the same as the pre-established pose categories given in the original dataset (e.g., looking left, right, upward, downward, and forward). No verbal behavioral labeling like attentive and distracted was done. Therefore, analysis of inter-rater reliability was not needed, as the classification task did not involve any subjective interpretation of behavior, but pose orientation classes.

In preprocessing, all images were down sampled to a 48×48 -pixel image to make the input size of all models

homogenous. The pictures were converted into grayscale after Red Green Blue (RGB) conversion to simplify the computation and normalize the input representation. The visual input to machine learning models was prepared as pixel intensities. This study does not use any handcrafted feature extraction technique like Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), or geometric landmark extraction. To the traditional classifiers, the scaled down grayscale images were converted into numerical input vectors. For the CNN model, the image matrices were used directly so that feature extraction could be learned automatically by the network during training. In this study, no explicit feature extraction or landmark detection (e.g., HOG, LBP, or facial landmark coordinates) was applied. Instead, the resized grayscale images were directly used as input data. For the conventional classifiers (SVM, DT, and RF), the 48×48 pixel images were flattened into numerical feature vectors representing pixel intensity values. For the CNN model, the image matrices were used directly so that the network could automatically learn relevant visual features during training. The distribution of image samples across the pose categories is shown in Fig. 2.

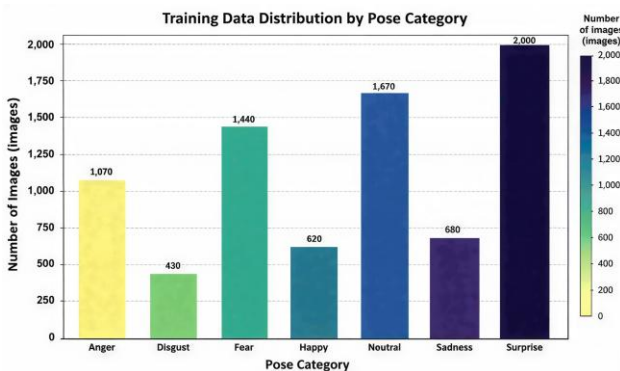


Fig. 2. Distribution of image samples across pose categories in the dataset.

B. Tools

Python was used to conduct the experiments in Google Colab. Image processing and preprocessing were done using OpenCV. The training and evaluation procedures were run in Google Colab, which offered a convenient cloud-based system, with Python being used to implement the machine learning and deep learning models.

C. Algorithms

In this study, four classification models were evaluated: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and Convolutional Neural Network (CNN). These models were selected to compare conventional machine learning approaches with a deep learning approach under the same preprocessing, data-splitting, cross-validation, and evaluation conditions.

1) SVM

Model overview: SVM is a supervised classification algorithm that identifies a decision boundary with a maximum margin and is effective for high-dimensional feature spaces.

Application in this study: Each preprocessed 48×48 grayscale image was flattened into a pixel-intensity feature vector and used to classify the corresponding pose-orientation category.

Key parameter settings and training method: Different kernel functions and regularization parameter (C) values were examined during 5-fold cross-validation on the training set, after which the selected configuration was retrained using the complete training subset.

Rationale for model selection: SVM was included because it provides a strong conventional baseline for image classification when the number of features is large relative to the dataset size.

2) DT

Model overview: DT is an interpretable supervised learning algorithm that recursively partitions the feature space into decision regions according to selected feature values.

Application in this study: The classifier received the flattened 48×48 grayscale pixel-intensity vectors and predicted the corresponding pose-orientation class.

Key parameter settings and training method: Entropy was used as the split criterion, the maximum tree depth was set to 10, and the remaining parameters retained their default values; model selection was performed using 5-fold cross-validation on the training set.

Rationale for model selection: DT was included as a transparent baseline whose decision process can be examined more easily than that of more complex models.

3) RF

Model overview: RF is an ensemble classifier that combines multiple decision trees built from bootstrap samples and randomly selected feature subsets, improving robustness and reducing the risk of overfitting compared with a single tree.

Application in this study: The model used the flattened 48×48 grayscale pixel-intensity vectors to classify the pose-orientation categories.

Key parameter settings and training method: The number of estimators and related RF hyperparameters were optimized through randomized search with 5-fold cross-validation on the training set, and the selected model was then retrained on the full training subset.

Rationale for model selection: RF was included to assess whether an ensemble of trees could provide more stable classification than an individual DT.

4) CNN

Model overview: CNN is a deep learning architecture that learns hierarchical spatial features directly from image data and is widely used for visual classification tasks.

Application in this study: The preprocessed 48×48 grayscale image matrices were supplied directly to the network to classify the pose-orientation categories without handcrafted feature extraction.

Key parameter settings and training method: The architecture contained three convolutional blocks, each comprising two convolutional layers followed by one max-pooling layer; Rectified Linear Unit (ReLU) activation was applied, and stochastic gradient descent was

used for optimization. The model was trained using the same training subset and validation procedure as the other classifiers.

Rationale for model selection: CNN was included to compare automatically learned image features with the flattened pixel representations used by SVM, DT, and RF.

Pre-trained architectures such as VGG-16 and ResNet were not used as baseline models because the objective was to compare SVM, DT, RF, and a CNN trained directly on the same dataset and preprocessing pipeline. Transfer-learning models will be considered in future work.

IV. RESULTS AND DISCUSSION

The performance of the evaluated models was assessed using accuracy, precision, recall, F1-Score, and confusion matrices. The primary metric used in the comparison of overall classification performance of the models were accuracy and precision; recall and F1-Score were more detailed class-level measures. Misclassification patterns between pose categories were also analyzed by studying the confusion matrices.

The overall comparison of the accuracy of the assessed models is given in Fig. 3. Among the four tested classifiers, SVM achieved the highest classification accuracy of 95%, followed by Decision Tree with 89%, Random Forest with 85%, and CNN with 72%. These results indicate that SVM was the most effective model for the considered pose-orientation classification task on the employed dataset.

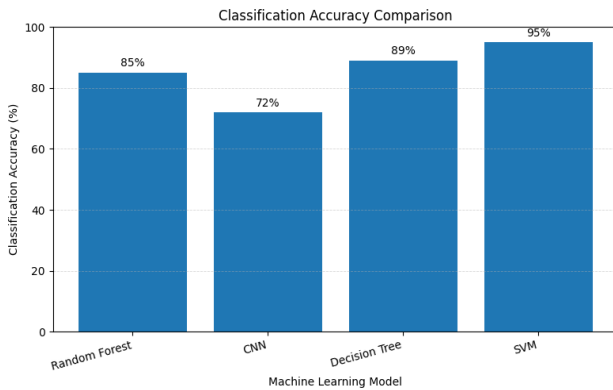


Fig. 3. Accuracy comparison of the evaluated machine learning models (SVM, DT, RF, and CNN).

To provide a standardized and clear comparison of the evaluated models, the performance metrics, including accuracy, precision, recall, and Macro F1-Score, are summarized in Table I.

As shown in Table I, the SVM model achieved the highest performance across all evaluation metrics. DT and RF also demonstrated strong and comparable results, while the CNN model showed relatively lower performance on the current dataset. The inclusion of Macro F1-Score provides a balanced evaluation across all classes, confirming the robustness of the SVM model for this classification task.

TABLE I. STANDARDIZED PERFORMANCE COMPARISON OF MACHINE LEARNING MODELS

Model	Accuracy (%)	Precision	Recall	Macro F1-Score
SVM	95	0.95	0.92	0.93
Decision Tree	89	0.87	0.86	0.86
Random Forest	85	0.84	0.86	0.85
CNN	72	0.70	0.68	0.63

The four considered machine learning models (SVM, DT, RF, and CNN), were tested on the pose-orientation image dataset using the identical preprocessing and evaluation environment. The SVM model has the best overall classification accuracy of 95% as demonstrated in Fig. 3. The accuracy of the Decision Tree model was 89%, and that of the Random Forest and CNN were 85% and 72%, respectively. These results indicate the SVM classifier was the most appropriate model among the models tested to handle this pose-based classification problem.

True positives and true negatives are accurately predicted perceptions. We need to keep false positives and false negatives to a minimum so that they don't show up in red shading. These terms are a little perplexing. So, let's go over each term one by one and make sure we understand it completely. The general structure and interpretation of these outcomes are illustrated in Fig. 4.

	PREDICTED p^+	PREDICTED n^+
ACTUAL p	True Positive	False Negative
ACTUAL n	False Positive	True Negative

Fig. 4. Structure of the confusion matrix used for model performance evaluation.

A confusion matrix was used to further analyze classification performance by showing the number of correctly and incorrectly predicted samples for each class. This helped identify which pose categories were most frequently confused by each model.

The detailed class-level performance of the RF model is presented in Figs. 5 and 6. Similarly, Figs. 7 and 8 show the DT results, Figs. 9 and 10 show the CNN results, and Figs. 11 and 12 show the SVM results.

	precision	recall	f1-score	support
0	0.75	0.86	0.80	72
1	0.86	0.85	0.85	112
2	0.91	0.86	0.75	72
3	0.96	0.88	0.89	123
4	0.87	0.86	0.92	106
5	1.00	0.92	0.86	56
6	0.69	0.88	0.77	21
accuracy			0.85	506
macro avg	0.84	0.86	0.85	506
weighted avg	0.87	0.86	0.86	506

Fig. 5. Classification performance metrics (precision, recall, and F1-Score) for the RF model.

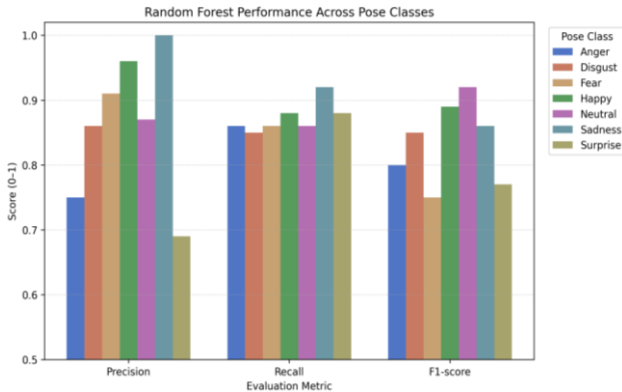


Fig. 6. Precision, recall, and F1-Score comparison for the RF model across pose classes.

As shown in Figs. 5 and 6, the RF model achieved generally balanced precision, recall, and F1-Scores across most pose classes, with macro-average values of 0.84, 0.86, and 0.85, respectively. However, performance varied among classes, indicating that some pose categories were more difficult to distinguish than others.

Figs. 7 and 8 indicate that the DT model produced stronger overall class-level results than RF, with macro-average precision, recall, and F1-Score values of 0.87, 0.86, and 0.86, respectively. Most classes were classified reliably, although a small number of categories showed comparatively lower precision or recall.

	precision	recall	f1-score	support
0	0.93	0.93	0.93	46
1	0.88	1.00	0.93	14
2	0.71	0.79	0.75	19
3	0.90	0.90	0.90	30
4	0.86	0.67	0.75	9
5	1.00	0.92	0.96	50
6	0.77	0.83	0.80	29
accuracy			0.89	197
macro avg	0.87	0.86	0.86	197
weighted avg	0.89	0.89	0.89	197

Fig. 7. Classification performance metrics (precision, recall, and F1-Score) for the DT model.

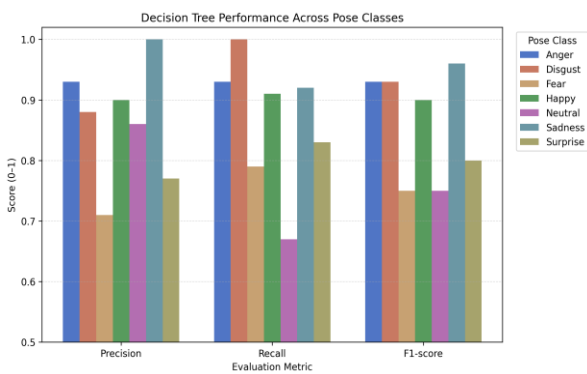


Fig. 8. Precision, recall, and F1-Score comparison for the DT model across pose classes.

The CNN results in Figs. 9 and 10 show more uneven performance across the pose classes. Its macro-average precision, recall, and F1-Score were approximately 0.70, 0.68, and 0.63, respectively, confirming that the CNN had

greater difficulty separating several visually similar categories in the current dataset.

	precision	recall	f1-score	support
0	0.8473	0.8222	0.8346	135
1	0.1869	1.0000	0.3149	54
2	0.9518	0.8927	0.9213	177
3	0.0000	0.0000	0.0000	75
4	1.0000	0.5845	0.7378	207
5	0.9216	0.5595	0.6963	84
6	0.9865	0.8835	0.9322	249
accuracy			0.7248	981
macro avg	0.6992	0.6775	0.6339	981
weighted avg	0.8390	0.7248	0.7503	981

Fig. 9. Classification performance metrics (precision, recall, and F1-Score) for the CNN model.

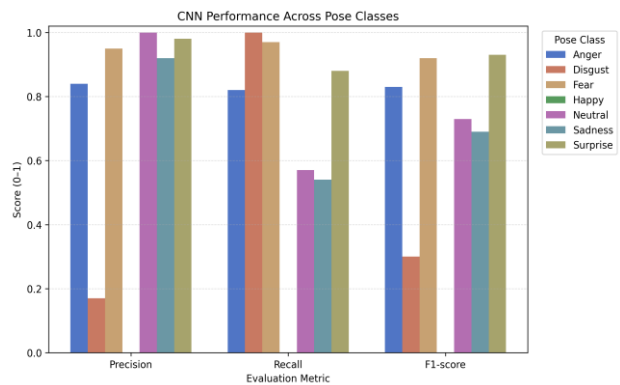


Fig. 10. Precision, recall, and F1-Score comparison for the CNN model across pose classes.

In contrast, Figs. 11 and 12 demonstrate that SVM achieved the most consistent class-level performance, with macro-average precision, recall, and F1-Score values of 0.95, 0.92, and 0.93, respectively. The high and relatively stable scores across the classes support the overall finding that SVM was the best-performing classifier in this study.

	precision	recall	f1-score	support
0	0.89	0.97	0.93	33
1	1.00	0.89	0.94	9
2	0.97	1.00	0.99	39
3	0.91	0.83	0.87	12
4	0.97	1.00	0.99	38
5	0.93	0.74	0.82	19
6	0.98	1.00	0.99	47
accuracy			0.95	197
macro avg	0.95	0.92	0.93	197
weighted avg	0.95	0.95	0.95	197

Fig. 11. Classification performance metrics (precision, recall, and F1-Score) for the SVM model.



Fig. 12. Precision, recall, and F1-Score comparison for the SVM model across pose classes.

Fig. 13 presents the confusion matrices of the four classifiers that were evaluated. These matrices give a closer look at the properly and improperly classified samples in each pose category. Of the models that were tested, the SVM classifier created the lowest amount of misclassification and this is expected as it had the highest overall accuracy. This finding also supports the fact that SVM was the best classifier in the given pose-orientation classification problem.

Even though the current experiment was performed as an offline image classification experiment, the suggested

methodology could be applied to real-time online learning systems in the future. Computer vision techniques may be used in a real-world context, where the images obtained by webcams during live online classes are analyzed to determine the head position and other visual signals. The results of the trained classifier would then be incorporated into a monitoring dashboard that would provide instructors with aggregate information about student attentiveness in online lectures.

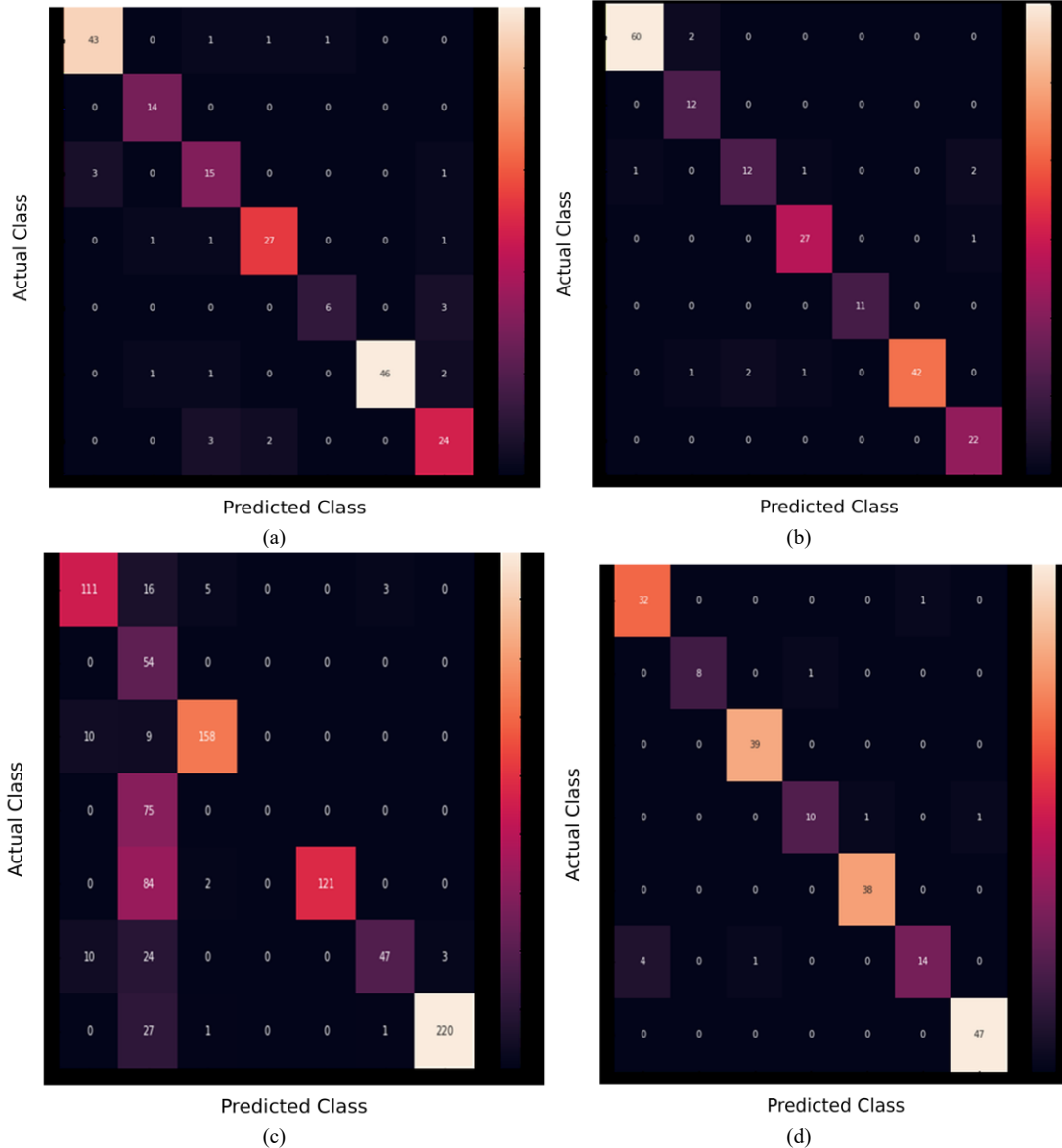


Fig. 13. Confusion matrix for the pose classification model. (a) RF; (b) DT; (c) CNN; (d) SVM.

V. CONCLUSION

This paper investigated the use of machine learning methods to classify visual cues related to body orientation that may be useful for future analysis of student attentiveness in online learning environments. A publicly available pose-orientation image dataset from Kaggle was

used to evaluate four classification models: SVM, DT, RF, and CNN.

The experimental results showed that SVM achieved the highest classification accuracy of 95%, followed by DT with 89%, RF with 85%, and CNN with 72%. These findings indicate that machine learning techniques, particularly SVM, can be effective for classifying pose-related visual patterns from image data.

One of the main limitations of this study is that the dataset does not consist of actual online classroom recordings. Therefore, the current findings should be interpreted as an initial proof of concept rather than a complete online student behavior prediction system. Future work should focus on collecting real online classroom data and incorporating additional behavioral cues, such as gaze direction, facial expression, and posture dynamics, to improve prediction accuracy and support real-time engagement analysis.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research was funded by the Arab Open University–Oman, Research Office under Internal Grant (Fund No.: BFP/GRG/ICT/25/185).

ACKNOWLEDGMENT

The authors thank Arab Open University–Oman for support.

REFERENCES

- [1] D. Mehta, M. F. H. Siddiqui, and A. Y. Javaid, “Facial emotion recognition: A survey and realworld user experiences in mixed reality,” *Sensors*, vol. 18, no. 2, 416, 2018.
- [2] M. Elbawab and R. Henriques, “Machine learning applied to student attentiveness detection: Using emotional and non-emotional measures,” *Education and Information Technologies*, vol. 28, no. 12, pp. 15717–15737, 2023.
- [3] A. De Giusti, “Policy brief: Education during COVID-19 and beyond,” *Revista Iberoamericana de Tecnología en Educación y Educación en Tecnología*, no. 26, pp. 110–111, 2020. doi: 10.24215/18509959.26.e12
- [4] M. Abbas and S. M. Khan, “Nonverbal communication and affectivity: A scale for detecting emotions through nonverbal behavior,” *Psychological Studies*, vol. 68, no. 3, pp. 472–480, 2023.
- [5] S. Sharma and P. Vyas, “Enhancing non-verbal communication in online classes: a conceptual framework,” *Journal of Education for Teaching*, vol. 48, no. 1, pp. 135–137, 2022.
- [6] Z. Zhang, Z. Li, H. Liu, T. Cao, and S. Liu, “Data-driven online learning engagement detection via facial expression and mouse behavior recognition technology,” *Journal of Educational Computing Research*, vol. 58, no. 1, pp. 63–86, 2020.
- [7] N. Yan and O. T. S. Au, “Online learning behavior analysis based on machine learning,” *Asian Association of Open Universities Journal*, vol. 14, no. 2, pp. 97–106, 2019. doi: 10.1108/AAOUJ-08-2019-0029
- [8] B. Fasel and J. Luettin, “Automatic facial expression analysis: a survey,” *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.
- [9] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, “Enhancing CNN with preprocessing stage in automatic emotion recognition,” *Procedia Computer Science*, vol. 116, pp. 523–529, 2017.
- [10] I. M. Revina and W. R. S. Emmanuel, “Face expression recognition using LDN and dominant gradient local ternary pattern descriptors,” *Journal of King Saud University–Information Sciences*, vol. 33, no. 4, pp. 392–398, 2021.
- [11] A. Acharya and D. Sinha, “Early prediction of students performance using machine learning techniques,” *International Journal of Computer Applications*, vol. 107, no. 1, pp. 37–43, 2014.
- [12] D. Ngo, A. Nguyen, B. Dang, and H. Ngo, “Facial expression recognition for examining emotional regulation in synchronous online collaborative learning,” *International Journal of Artificial Intelligence in Education*, vol. 34, no. 3, pp. 650–669, 2024.
- [13] C. Pabba and P. Kumar, “A vision-based multi-cues approach for individual students’ and overall class engagement monitoring in smart classroom environments,” *Multimedia Tools and Applications*, vol. 83, no. 17, pp. 52621–52652, 2024.
- [14] I. Lasri, A. Riadsolh, and M. Elbelkacemi, “Facial emotion recognition of deaf and hard-of-hearing students for engagement detection using deep learning,” *Education and Information Technologies*, vol. 28, no. 4, pp. 4069–4092, 2023.
- [15] L. V. L. Pinto *et al.*, “A systematic review of facial expression detection methods,” *IEEE Access*, vol. 11, pp. 61881–61891, 2023.
- [16] A. A. Alhussan *et al.*, “Facial expression recognition model depending on optimized support vector machine,” *Computers, Materials & Continua*, vol. 76, no. 1, 2023.
- [17] C. Bisogni, L. Cimmino, M. De Marsico, F. Hao, and F. Narducci, “Emotion recognition at a distance: The robustness of machine learning based on hand-crafted facial features vs deep learning models,” *Image and Vision Computing*, vol. 136, 104724, 2023.
- [18] M. Irfan, P. Patel, and B. Hassan, “EngageSense: a hybrid approach for real-time engagement detection for virtual classrooms,” in *Proc. 2025 IEEE Global Engineering Education Conference*, 2025, pp. 1–9.
- [19] R. K. Vani and P. Jayashree, “Multimodal emotion recognition system for e-learning platform,” *Education and Information Technologies*, vol. 30, no. 10, pp. 13507–13538, 2025.
- [20] A. M. Mutawa and S. Sruthi, “Enhancing human–computer interaction in online education: a machine learning approach to predicting student emotion and satisfaction,” *International Journal of Human-Computer Interaction*, vol. 40, no. 24, pp. 8827–8843, 2024.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).