computers

MDPI

Article

Transformer-Based Student Engagement Recognition Using Few-Shot Learning

Wejdan Alarefah *, Salma Kammoun Jarraya * and Nihal Abuzinadah

Computer Science Department, Faculty of Computing and Information Technology, King Abdulaziz University (KAU), Jeddah 21589, Saudi Arabia; nabuznadah@kau.edu.sa
* Correspondence: wabdullahalghamdi0002@stu.kau.edu.sa (W.A.); smohamad1@kau.edu.sa (S.K.J.)

Abstract: Improving the recognition of online learning engagement is a critical issue in educational information technology, due to the complexities of student behavior and varying assessment standards. Additionally, the scarcity of publicly available datasets for engagement recognition exacerbates this challenge. The majority of existing methods for detecting student engagement necessitate significant amounts of annotated data to capture variations in behaviors and interaction patterns. To address these limitations, we investigate few-shot learning (FSL) techniques to reduce the dependency on extensive training data. Transformer-based models have shown comprehensive results for video-based facial recognition tasks, thus paving new ground for understanding complicated patterns. In this research, we propose an innovative FSL model that employs a prototypical network with the vision transformer (ViT) model pre-trained on a face recognition dataset (e.g., MS1MV2) for spatial feature extraction, followed by an LSTM layer for temporal feature extraction. This approach effectively addresses the challenges of limited labeled data in engagement recognition. Our proposed approach achieves state-of-the-art performance on the EngageNet dataset, demonstrating its efficacy and potential in advancing engagement recognition research.

Keywords: few-shot learning; vision transformer; student engagement recognition

Academic Editors: M. Ali Akber Dewan and Selene Tomassini

Received: 11 February 2025
Revised: 7 March 2025
Accepted: 8 March 2025
Published: 18 March 2025

Citation: Alarefah, W.; Jarraya, S.K.; Abuzinadah, N. Transformer-Based Student Engagement Recognition Using Few-Shot Learning. *Computers* **2025**, *14*, 109. https://doi.org/10.3390/computers14030109

1. Introduction

The use of online learning has become mainstream, and it has recently played a key role in the educational field. Online learning helps students by taking advantage of computer techniques, allowing teachers to provide lessons to students efficiently [1].

However, the recognition of e-learning engagement is a critical issue in educational technology. Traditional methodologies are insufficient to assess engagement in all situations, and student performance in online learning environments often suffers due to the limited interaction between teachers and students [1]. In addition, due to the complexities of student participation, as well as the influence of diverse definitions and standards, its evaluation and measurement are also challenging. As a result, teachers are unable to assess the level of student engagement; thus, methods for the efficient and automatic recognition of students' learning engagement are required.

Typical approaches for identifying and assessing engagement include (a) self-reporting, (b) observational checklists and rating scales, and (c) automated measures that use technical tools to detect students' engagement, such as facial expression recognition [2] [3,4], body gesture recognition [5], and head pose and eye gaze tracking [6].

The automated measurements are more objective than the other two methods [3]. Most of these measurements come from computer vision-based tools which have been

shown to be effective in detecting e-learners' degrees of engagement [5]. Computer vision methodologies have demonstrated greater appropriateness for online learning due to their reduced distractibility for users, alongside the widespread availability and affordability of the requisite equipment and software for data recording and evaluation [7]. One of the biggest challenges in this field is the extremely limited number of datasets relating to student engagement. The majority of existing methods for detecting student engagement necessitate significant amounts of annotated data to capture variations in behaviors and interaction patterns, which can be expensive at times [8]. Moreover, traditional supervised learning methods demonstrated limited performance in the student engagement recognition task. However, few-shot learning techniques can be implemented in order to reduce the amount of data used/required in training. Although few-shot learning approaches have significantly advanced computer vision, to the best of our knowledge, they have not yet been explored in the context of student engagement recognition in online learning.

In this research, we aim to investigate the appropriateness of few-shot learning for student engagement recognition using the dataset produced by [5] and the EngageNet dataset [9].

Few-shot learning (FSL) is a meta-learning problem in which models are evaluated through an N-way, K-shot classification problem [10], in which the model learns k samples from N classes. The few-shot learning technique can identify novel classes using only a few samples once it has been deployed [8]. The objective of FSL is to overcome the obstacles faced by deep learning techniques, which include the rarity of samples, the high effort required for collecting data, and the high cost of the computational process [8]. FSL is a step on the way to mimicking human-like learning. Hence, FSL has been used in a wide range of real-world applications, including computer vision, robotics, acoustic signal processing, and natural language processing [8].

Vision Transformers (ViTs), as proposed in [11], have become a powerful feature representation model and were recently used widely, including the following research works: [11–13]. Vision Transformers provide comparable performance for understanding video-based facial recognition context.

In this research, we propose an FSL model that employs a prototypical network with the vision transformer (ViT) model pre-trained on a face recognition dataset (e.g., MS1MV2) for spatial feature extraction, followed by an LSTM layer for temporal feature extraction. We investigated the suitability of few-shot learning for recognition of student engagement level and explored the accuracy achieved through utilizing the specific architecture on the dataset collected in [5]. The novelty of our approach lies in the combination of Vision Transformers and Few-Shot Learning to handle the challenges of limited labeled data in engagement recognition, unlike traditional ML methods which rely hugely on the amount of training data.

The rest of this paper is organized as follows: Section 2 reviews the related work on student engagement recognition, few-shot learning, and Vision Transformers. Section 3 details the proposed methodology. Section 4 presents the experimental setup and results, including the dataset and model architecture, and Section 5 concludes with a discussion and future directions.

2. Literature Review

2.1. Student Engagement Recognition

In the context of online learning, many uncontrollable factors affect student engagement, such as the learning environment and information interruption; therefore, teachers must use a system that can recognize student engagement during online learning [1]. Various studies have explored student engagement recognition in online learning, ranging from

single-model to multi-model approaches [14]. Visual cues are used in computer vision-based approaches, such as facial expressions [2–4], body gestures [5], and eye gaze [6].

Zhang et al. [3] used mouse movements from students' facial expressions to improve labeling accuracy. Later on, adaptive weighted Local Gray Code Patterns and quick sparse representation techniques were employed for feature extraction and classification. Altuwairqi et al. [2] carried out a number of investigations into the recognition of students' engagement levels depending on their emotions; they linked each level of engagement with specific emotions by computing the Matching-Score (MS) and Mis-Matching Score (MisMS) for both matched and unmatched emotions at each engagement level.

The usability of convolutional neural networks (CNNs) has been investigated in this field. Nezami et al. [4] proposed a CNN model that was adapted from the VGG-B framework, and then they pre-trained the model on the FER-2013 dataset and fine-tuned it using their engagement recognition dataset (ER). Khenkar et al. [5] investigated a deep three-dimensional CNN model for the recognition of e-learners' engagements based on spatio-temporal features of micro-body gestures. The authors also used a transfer learning approach to the 3D CNN model trained on the Sports-1M dataset. The resulting accuracy shows the efficiency of using body gestures for engagement recognition. Another work [7] used multiple CNN architectures, including All-Convolutional-Network (ALL-CNN), Network-in Network (NiN-CNN), and Very-Deep-Convolutional-Network (VD-CNN). In addition, motivated by these models, the authors proposed a model that achieved higher accuracy for the students' engagement classification. Kaur et al. [6] used a Multi-Instance Learning (MIL) deep network for the prediction and localization of the learners' eye gaze movements and head pose characteristics, then passed these features through the LSTM-based network and flattened the output. The resulting vector passed through three dense layers and average pooling, producing a single regressed engagement value.

While Hasnine et al. [14] associated student emotions with engagement levels, their approach lacked validation with diverse datasets, which we address using few-shot learning techniques. The developed model consists of four phases conducted in the following order: the first phase involves face recognition using the OpenCV Library, emotion detection using CNN, eye detection, and engagement recognition using the Concentration index. The model was tested using videos captured from a web camera available on YouTube, including eleven students; therefore, the validation of the model did not use an appropriate dataset. While numerous machine learning and deep learning methodologies have been explored for the recognition of student engagement, few-shot learning techniques have not been thoroughly investigated for this purpose, to the best of our knowledge.

2.2. Few-Shot Learning Technique

Most deep learning methods are unable to learn from a few examples in real-world scenarios where data are scarce, and they tend to overfit. Therefore, there is a large volume of published studies describing the role of using only a few samples. A novel paradigm shift known as few-shot learning allows for the development of models that can rapidly learn a new category from a small number of training examples. Supervised machine learning models, on the other hand, need a significant volume of data to be more accurate. In few-shot learning, the model is trained using numerous training tasks (also called episodes) [8]. Each task contains its own support set and query set that include N classes and K samples (known as N-way, K-shot), as shown in Figure 1, and few-shot learning is called "one-shot learning" when there is only one sample per class. In each task, the model is trained on the support set and verified using the query set, after which the model is evaluated using a test task that contains its own support and query sets that are not included in the training [15]. In other words, once the few-shot learning model has been

trained, it will be able to classify new classes using previously acquired information and with the help of additional information (the support set) [16].

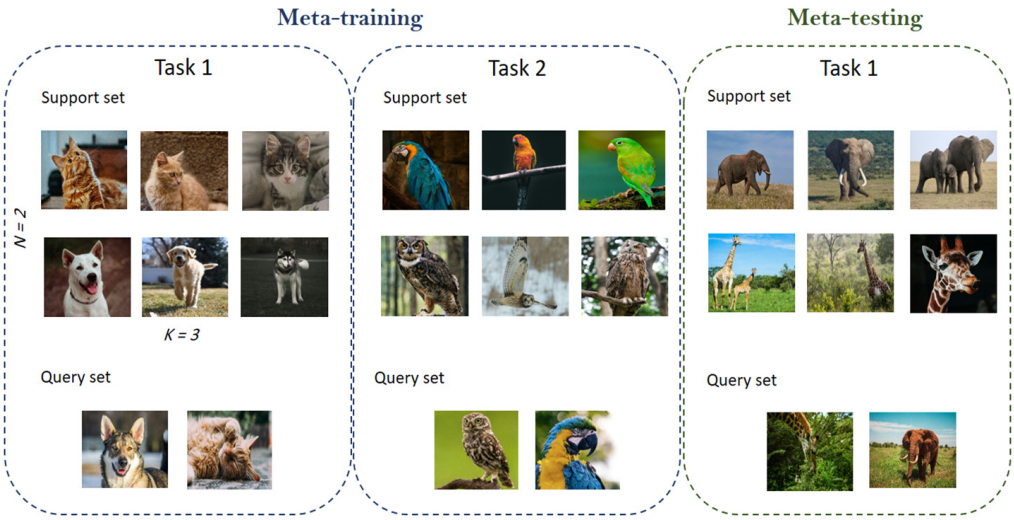

Figure 1. An example of a 2-way 3-shots classification task for few-shot learning.

According to Liu et al. [10], few-shot learning approaches based on meta-learning can be categorized into metric-, optimization-, and model-based learning.

Metric-based learning mainly consists of three phases, as follows. The initial step is to derive a metric space from a set of training samples using a network in which samples from the same class are near together and samples from other classes are far apart [17]. After the network has been trained, it can be regarded as an embedding function. The second phase is to extract features from all the testing data set samples. The final phase involves classifying the testing samples using similarity function (Euclidean or Cosine distance) [18].

Maddula et al. [17] introduced a Meta-Learning Approach to Recognize Emotions (MLARE), utilizing a Siamese network alongside a binary cross-entropy loss mechanism (BCE) combined with sigmoid activation as the loss function in order to improve accuracy. The Prototypical Network [15] is another metric-based learning approach that uses a non-linear neural network which maps the input into an embedding space and defines the prototype of each class as the average of its support set within the embedding space. The classification of an embedded query point is subsequently executed by locating the closest class prototype.

Sung et al. [19] proposed Relation Network, a flexible metric-based model comprised of two modules; first, the used embedding module concatenates the resulting feature vector of each sample in the support set with the feature vector of the query sample, then feeds them into the relation module, which calculates the relation score between the query sample and each sample in the support set.

2.3. Vision Transformer

Transformers' uncomplicated architecture enables the processing of a variety of modalities (e.g., images, videos, text, and audio) with comparable processing blocks. Additionally, it has the capability to efficiently replace the CNN models in deep neural networks since the pioneering development of the Vision Transformer (ViT) [11], which introduced the use of transformers for image classification tasks with minimal changes by dividing each image into patches, embedding them, and concatenating the embeddings with positional encodings before passing them to the transformer block. As illustrated in Figure 2, the transformer block comprises a multi-head attention layer and a multi-layer perceptron (MLP) layer, each preceded by normalization layers. Dosovitskiy et al. [11] trained the

transformers on very large dataset, revealing that data-hungry models. Then, Touvron et al. [20] produced the Data-efficient Image Transformer (DeiT) to demonstrate that transformers can be trained with mid-size datasets (e.g., ImageNet-1k) by leveraging several data augmentation methods and novel distillation techniques. The research [12] explored that the ViT can perform well on smaller datasets in the field of face recognition through employing patch-level data augmentation techniques.

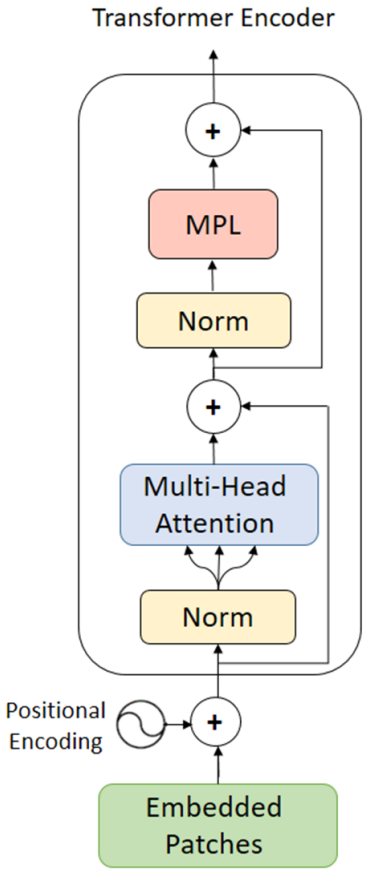

Figure 2. The architecture of the Vision Transformer Encoder.

Some studies have investigated the use of transformers for recognizing student engagement [13,16,21]. However, as previously stated, a major challenge in this field is the extremely limited availability of data for recognizing student engagement levels.

Therefore, this research investigates the efficiency of applying few-shot learning techniques to improve online learning methods. To the best of our knowledge, combining Vision Transformers (ViTs) with LSTM layers in few-shot learning scenarios remains a relatively unexplored area of research.

3. Proposed Architecture

The proposed few-shot learning model (see Figure 3) integrates a Vision Transformer (ViT) with a Long Short-Term Memory (LSTM) network to extract spatio-temporal features for engagement recognition in videos of students learning online. Additionally, it leverages the episodic training approach with prototypical loss to improve the generalization capability when using limited labeled data. The first stage involves preprocessing videos by extracting 16 frames from each, followed by normalization and resizing to 112 $\times$ 112 pixels. The second stage involves the feature extraction model, and the last stage includes computing the prototypes and the losses.

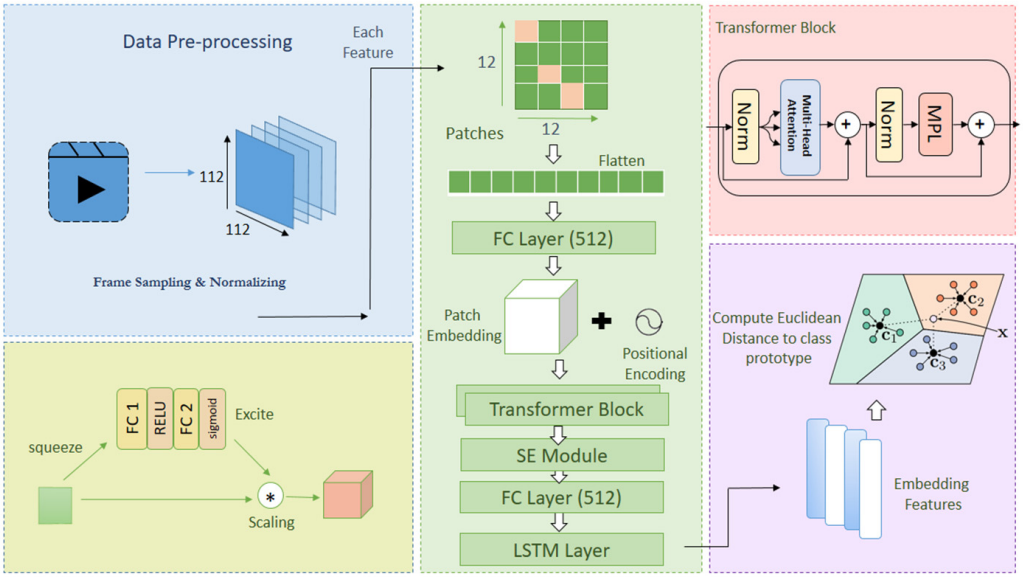

Figure 3. The overall architecture of our approach. We utilized the vision transformer model with the SE module to enhance the extracted features, followed by a dimensionality reduction layer and LSTM layer for temporal feature extraction; then, the prototypes were computed for classification.

Vision Transformer Backbone

We selected the TransFace ViT model [12] pre-trained on the MS1MV2 dataset, due to its proven performance in facial recognition tasks, which are critical for detecting engagement cues. Figure 3 shows how we integrate the TransFace model with the LSTM layer.

The Vision Transformer (ViT) processes each frame independently to extract spatial features. We follow the typical Transformer formulation. Key operations in this module include patch embedding, positional encoding, and transformer-based feature extraction. Each frame is divided into P × P non-overlapping patches; then, each patch is flattened and projected into an embedding space of dimension D. To retain spatial information, a learnable positional encoding is added to the patch embeddings.

A stack of transformer blocks processes the token sequence. Each block consists of normalization layers, a multi-head self-attention mechanism and a feed-forward network (FFN).

$$H = A(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

The self-attention mechanism A enhances the dot-product attention operation by organizing three components: queries, keys, and values (q,k,v) into matrix structures (Q,K,V). The mechanism for self-attention utilizes a softmax function to determine the weights of attention for each value in V by calculating a dot product among all queries in Q and all keys in K.

The output of the ViT for each frame is a high-dimensional feature vector, which is flattened into a 1D representation for further processing. Next, each feature vector is scaled using the squeeze and excitation module which highly improve the accuracy. Then, to reduce the computational complexity and align the feature dimensions with the LSTM input requirements, a linear layer is applied. The LSTM network processes sequences of feature vectors to capture temporal dynamics.

The model is trained using prototypical loss in an episodic manner of training. For each class c, a prototype P_c is computed as the mean of the support embeddings. We use

Euclidean distance as a similarity function, and then the probability of the query be-longing to class c is given by the softmax over distances, as follows:

$$P(y = c|f_{query}) = \frac{exp\ (-d(f_{query}, P_c))}{\sum_{\acute{c}} exp\ (-d(f_{query}, P_{\acute{c}}))} \quad (2)$$

The prototypical loss minimizes the negative log-likelihood of the correct class:

$$L_{proto} = -\frac{1}{n_{query}} \sum_{i=1}^{n_{query}} log\ P(y_i = c_i|f_{query,i}) \quad (3)$$

4. Experimental Results

4.1. Dataset

To evaluate our approach, we employed two student engagement recognition datasets: the Khenkar Dataset [5] for training and the EngageNet dataset [9] for testing to follow the few-shot learning settings.

Khenkar Dataset [4]: The datasets in the field of student engagement recognition are very limited, and one of the most reliable available datasets is the Khenkar Dataset [5], which contains multi-class videos (High-, Medium-, Low-Engagement and Disengagement) annotated by experts using the emotion-based affective model [2]. Over 2476 video clips are included in the dataset. Each video clip ranges from 2 to 40 s long and was collected from 24 lectures involving five college students. The dataset was collected in an uncontrolled setting and captured using built-in webcams, representative of the natural environment of an e-learning student. We used the color jetter and horizontal flip augmentation techniques to balance the unbalanced classes in the dataset. The final class distribution is shown in Table 1. We then split the dataset as 70% for training and 30% for validation.

Table 1. The sample distribution on the Khenkar Dataset.

Engagement Level	# of Samples
High engagement	2936
Medium engagement	2199
Low engagement	1890
Disengagement	1090

EnagageNet dataset [9]: The EngageNet dataset was utilized in the meta-testing stage of our approach. Each clip contains 10 s of video at a frame rate of 30 frames per second and a size of 1280 × 720 pixels. Four levels of engagement have been assigned to the video records of the subjects: "Not Engaged", "Barely Engaged", "Engaged", and "Highly Engaged". A subject-independent data split method was used to divide the dataset into 7983 samples for training, 1071 samples for validation, and 2257 samples for testing [9]. However, the testing data are not available to the public, so we used the validation set to test our model. During the experiments, we observed a high overlap between classes from the EngageNet dataset. We used Maximum Mean Discrepancy (*MMD*) [22], which represents the distance between distributions as follows.

$$MMD^2 = \frac{\sum_{i \neq j} K(X_i, X_j)}{n(n-1)} + \frac{\sum_{i \neq j} K(Y_i, Y_j)}{m(m-1)} - 2 \cdot \frac{\sum_{i,j} K(X_i, Y_j)}{n \cdot m} \quad (4)$$

The results of computing the *MMD* between the highly engaged class with the engaged and barely engaged classes were 0.003 and 0.01, respectively. This indicates a very small difference between their distributions. Therefore, we combined these three labels

"Highly-engaged", "Engaged", "Barely-engaged", into a single label, "Engaged", while the latter label, "Not-Engaged", remained unchanged. This then yielded a binary classification task, as outlined in Table 2. The testing process was episodic, and we selected relatively equal samples from each class.

Table 2. Sample distribution among the combined classes in the EnagageNet dataset.

Before	After	# of Samples
Highly-Engaged Engaged Barely-Engaged	Engaged	130
Not-Engaged	Not-Engaged	130

4.2. Implementation Details

In this research, we incorporated the TransFace pre-trained model [12] in our framework as a spatial feature extractor with an LSTM network for temporal features. We first split the video clips into smaller clips ranging from 3 to 10 s in length, and the number of samples in each class is shown in Table 1. Then, we used uniform sampling for frame extraction, which involves selecting 16 frames at regular intervals from a video and then normalizing and resizing each frame to be 112×112, which is the expected shape for the ViT model. We trained our model for 100 epochs with 20 episodes. Adam was chosen as the optimizer with a learning rate of 0.0001 and 0.001 as a weight decay. To optimize the learning process and prevent overfitting during training, we employed the ReduceLROnPlateau learning rate scheduler. This scheduler adaptively reduced the learning rate when validation loss plateaued, ensuring more efficient convergence.

The TransFace model was fine-tuned with 12 transformer blocks, each with eight attention heads; see Table 3.

Table 3. The details of the proposed model.

	Model
# of Frames/clip	16
Learning Rate	0.0001
Weight Decay	0.001
Epochs	100
Iterations/Episodes	20
Batch-Size	40
Transformer Blocks	12 each with 512 dim
Attention Heads	8
Optimizer	ReduceLROnPlateau
Loss	Prototypical Loss
Similarity Measure	Euclidean Distance

We trained the framework with four classes: high-engagement, medium-engagement, low-engagement, and disengagement classes. The proposed model achieved 97% training accuracy and 90% validation accuracy during the meta-training stage in a 4-way 5-shot scenario.

4.3. Evaluation Metrics

To assess our results we employed multiple performance measures including Accuracy, Precision, Recall, and F1-measure as follows:

$$\text{Accuracy} = \frac{\textit{True Positives} + \textit{True Negatives}}{\textit{Total Number of Samples}} \times 100\% \tag{5}$$

$$\text{Precision} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Positives}} \tag{6}$$

$$\text{Recall} = \frac{\textit{True Positives}}{\textit{True Positives} + \textit{False Negatives}} \tag{7}$$

$$\text{F1-measure} = 2 \times \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}} \tag{8}$$

These metrics were selected to provide a comprehensive evaluation of our model's performance in different aspects.

4.4. Results on the Unseen: EngageNet

Table 4 and Figure 4 present the testing results on the EngageNet dataset. Our results are obtained over 10 runs, each consisting of 26 iterations, where each iteration includes 5 support samples and 5 query samples. The proposed model achieved an overall accuracy of 73.62% ± 2.66% in the binary classification task, indicating a 95% confidence interval (CI) between 70.96% and 76.28%. According to the confusion matrix shown in Figure 4, the model correctly classified 75% (±4.9%) of Engaged instances (true positive rate) and 76% (±4.7%) of Not-Engaged instances (true negative rate). However, 25% of Engaged instances were misclassified as Not-Engaged, while 24% of Not-Engaged instances were misclassified as Engaged, suggesting challenges in differentiating subtle engagement cues.

Table 4. The few-shot evaluation with EngageNet dataset.

Metric	**2-Way 5-Shot**					
	TransFace ViT			**Swin ViT**		
	Engaged	**Not-Engaged**	**Avg.**	**Engaged**	**Not-Engaged**	**Avg.**
Precision	0.7495 ± 0.1153	0.7597 ± 0.1244	0.7546	0.5758 ± 0.15033	0.57615 ± 0.1412	0.5759
Recall	0.7251 ± 0.1527	0.7348 ± 0.1467	0.7299	0.5552 ± 0.1743	0.5819 ± 0.1740	0.5686
F1-Measure	0.7158 ± 0.1094	0.7235 ± 0.1077	0.7197	0.5445 ± 0.1385	0.5560 ± 0.1350	0.5503
Accuracy	0.7362 ± 0.0266			0.5686 ± 0.0231		

The confusion matrix reveals higher misclassification rates for engaged students, which is likely due to subtle differences between the 'Engaged' and 'Not-Engaged' classes that the model struggles to differentiate. Additionally, due to the limited number of samples in the 'Not-Engaged' class, some samples were repeated as part of the support set, while the 'Engaged' class contained more diverse samples.

As illustrated in Table 4 and Figure 5, the Not-Engaged class has higher recall (0.76) and a better F1 score (0.72) compared to the Engaged class (recall 0.74, F1 0.71). This indicates that the model is better at detecting "Not-Engaged" examples.

Furthermore, we evaluated the inference speed to assess the model's feasibility for real-time applications. We utilized an end device equipped with an Intel Core(TM) i7-7500U CPU (2.70 GHz) and 8 GB RAM, running python 3.10 with Pytorch 1.13. The proposed approach achieves an average inference time of 36.5705 s per image with a 95% confi-

dence interval of ±2.7354 s, indicating its potential for real-time engagement recognition in classrooms.

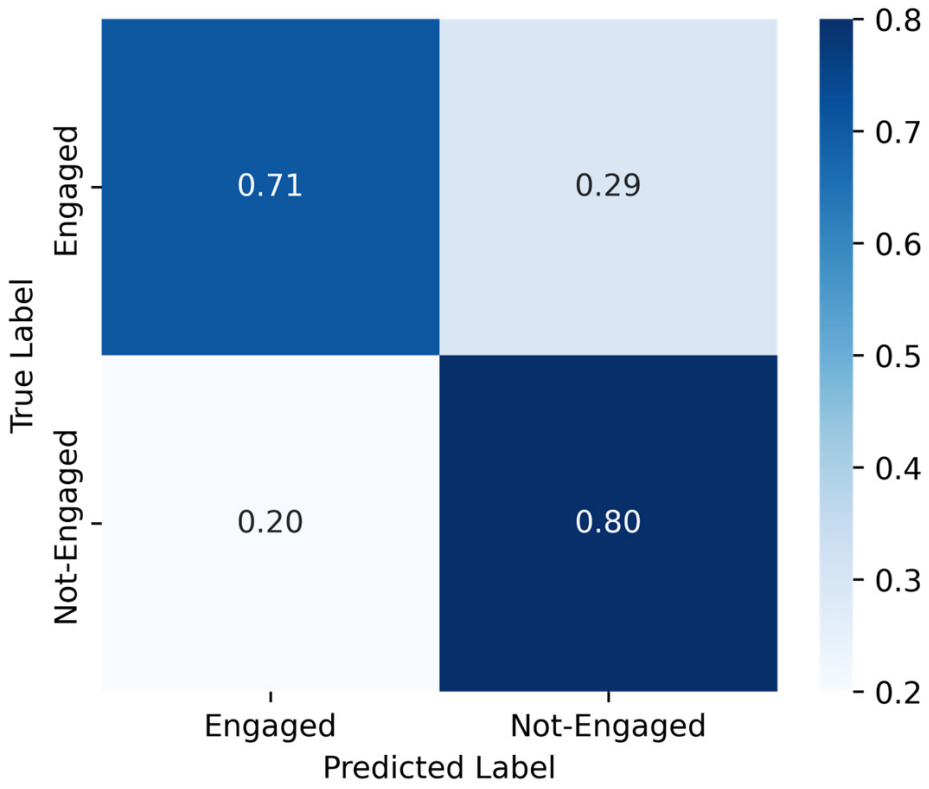

Figure 4. Confusion matrix of the proposed TransFace ViT + LSTM model on EngageNet dataset.

To evaluate our results, we explore our approach with the Swin ViT model [23], which is a benchmark backbone that demonstrated strong results in the field of image classification due to its robust hierarchical feature extraction capabilities, making it suitable for addressing challenges in engagement classification.

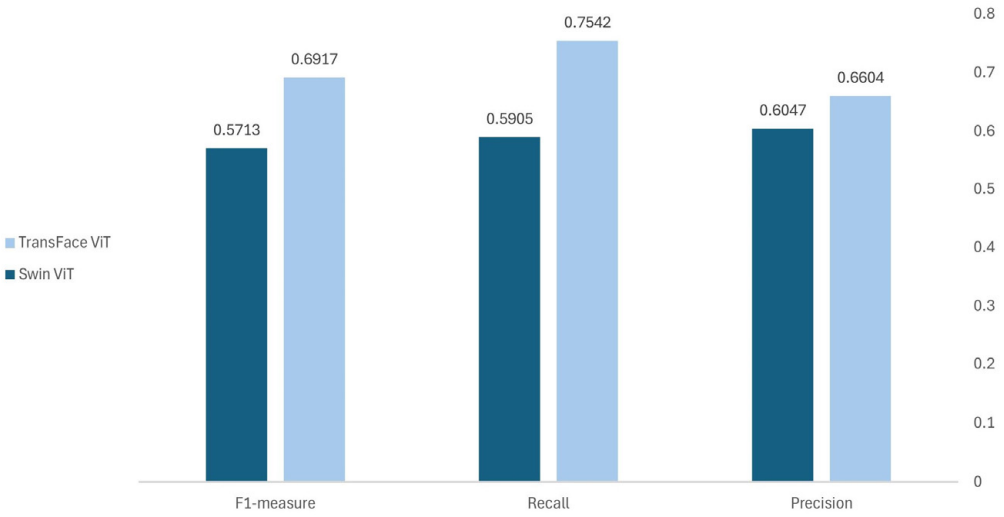

Figure 5. Comparison between the TransFace ViT model and the Swin ViT.

Under the same conditions, we combined the Swin ViT pre-trained model on the ImageNet-1K dataset with the LSTM layer. The model was fine-tuned alongside the remaining components using the same settings outlined in Table 2. The Swin ViT model achieved 98% training accuracy and 92% validation accuracy.

The TransFace ViT model outperforms the Swin ViT model with a significantly higher accuracy of 74% compared to 57%. This demonstrates the effectiveness of the TransFace ViT model in correctly classifying engagement levels. The results further confirm that the TransFace ViT model is better suited for binary classification tasks involving engagement recognition. While Swin ViT has shown success in image classification tasks, its performance in engagement recognition is comparatively weaker, particularly in terms of recall (0.57) compared to TransFace ViT's (0.74) and F1-measure (0.55) compared to TransFace ViT's (0.73). The overall performance of the proposed approach with the TransFace ViT model demonstrated consistent and promising results.

4.5. Comparison Performance

Table 5 compares various models that have been explored for student engagement recognition, utilizing diverse methods. The deep learning approaches, such as the EfficientNet B7 + LSTM model [24], and the transformer-based models, such as the Video Vision Transformer (ViViT) [25] and the Vision Transformer + Temporal Convolutional Network [16], have been implemented, and they, respectively, achieved 67.48%, 63.9%, and 65.58% improvements in temporal understanding. However, these methods face some challenges in distinguishing engagement levels. Another work employed the VGG16 fine-tuned model [26] and achieved 74.9% accuracy. However, traditional convolutional neural networks (CNNs) have been widely implemented but often require large amounts of labeled data. The Temporal Convolutional Network with Autoencoder (TCN-AE) [27] utilizes time-series data to capture behavioral and emotional cues, reporting an AUC ROC of 0.7489.

Table 5. Comparison of previous engagement measurement approaches with the proposed approach in this paper.

Ref	Feature Extraction	Method	Task Type	Accuracy
Mandia et al. [25]	Detect faces using the Multi-task Cascaded Convolutional Network (MTCNN)	Video Vision Transformer (ViViT) based architecture named Transformer Encoder with Low Complexity (TELC)	Multi-class	63.9%
Zhang et al. [16]	Facial features	Vision transformer + Temporal convolutional network	Multi-class	65.58
Selim et al. [24]	CNN	EfficientNet B7 + LSTM	Multi-class	67.48%
Abedi et al. [27]	Time-series data sequences extracted from both the behavioral feature and emotional states	Temporal convolutional network with autoencoder TCN-AE	Binary	(AUC ROC) 0.7489
Tieu et al. [26]	CNN	Fine-tune the VGG16	Binary	74.9%
(Proposed model)	Vision transformer	Proposed Model (ViT + LSTM + FSL)	Binary	74%

The proposed model (ViT + LSTM + few-shot learning) outperforms binary classification methods with 74% accuracy, demonstrating its effectiveness in student engagement recognition and addressing the challenges of limited labeled data. Integrating Vision Transformers for spatial feature extraction, LSTMs for temporal feature extraction, and few-shot learning to generalize from limited samples, the model enhances engagement recognition compared to traditional CNN- or LSTM-based methods. These findings suggest

that leveraging transformer architectures with few-shot learning can significantly improve engagement recognition in real-world educational settings.

5. Discussion

As the proposed few-shot learning model, which integrates a Vision Transformer (ViT) with an LSTM layer, was trained on a limited set of video clips and generalizes to unseen data with 75% accuracy, teachers can implement it in real classroom settings to automatically assess student engagement levels from video recordings captured via a webcam. This approach automates engagement recognition and helps educators adapt their teaching strategies based on students' responsiveness.

Our model showed lower classification rates on the Engaged students due to the challenges regarding the engagement recognition of the data collected in a realistic setting with diverse conditions and different participants' ages. The participants were free to move around and sometimes become far from their devices with no restrictions on lighting or backgrounds. These factors contributed to increased variability, making engagement recognition more complex.

To further investigate the results proposed by our model, we have utilized the t-SNE (t-distributed Stochastic Neighbor Embedding) model to visualize the feature space learned during the training. However, as shown in Figure 6, the t-SNE plot indicates that data points from different engagement levels (i.e., "Engaged" vs. "Not-Engaged") cluster closely in the reduced dimensional space, suggesting substantial overlap in their underlying features. Moreover, engagement is frequently a subtle, continuous cue rather than a precisely defined category variable. Our t-SNE results highlight that certain "borderline" samples share characteristics of both classes.

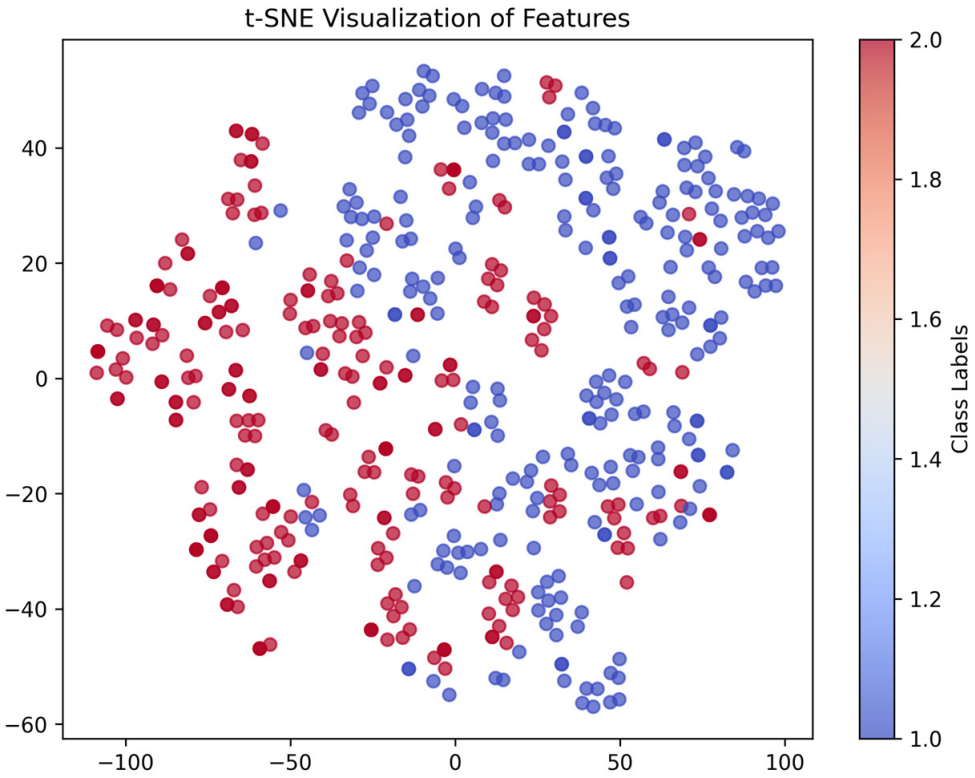

Figure 6. T-SNE visualization of extracted features from the EngageNet dataset.

Despite the model's promising performance, certain constraints about the utilized datasets must be noted. The Khenkar dataset, used for training, consists of only five students with video clips collected from their devices under varying lighting conditions

and camera angles. While this introduces some diversity, the small sample size may limit the model's ability to generalize to broader student populations. The EngageNet dataset, with 127 participants aged 18 to 37 years, provides a wider range of engagement expressions and was annotated by three expert observers. Both datasets include computer-based and in-the-wild settings and were annotated using behavioral (facial and body cues) and cognitive (self-reports) dimensions. However, a notable imbalance exists in the engagement labels, with the 'not engaged' class being significantly smaller than the other categories, limiting the number of test clips to 130. To improve generalizability, future work should explore the model's performance on larger and more diverse datasets, including different educational settings, age groups, and cultural backgrounds.

6. Conclusions and Future Directions

We introduced a novel approach for recognizing student engagement in online learning environments, integrating few-shot learning with Vision Transformers (ViT) and Long Short-Term Memory (LSTM) networks. The proposed model was combined with a prototypical loss in an episodic training approach to address the challenge of limited labeled data. The experimental results indicate that the proposed model achieved highly promising results on the EngageNet dataset. The results highlight the importance of incorporating specialized models such as TransFace ViT for engagement classification tasks.

Future research directions involve investigating real-time deployment scenarios and integrating additional few-shot learning techniques, such as optimization-based methods, to enhance model performance. Moreover, to enhance the practicality of the proposed model, future work should focus on improving inference speed and computational efficiency for real-time classroom applications. We intend to benchmark latency, memory usage, and energy consumption against existing models, addressing the trade-offs between accuracy and resource constraints. Additionally, future research should incorporate multimodal data (audio features and physiological signals). Integrating these modalities with the vision cue data will significantly enhance the model's performance for recognizing subtle features of engagement.

Author Contributions: Conceptualization, W.A. and S.K.J.; data curation, W.A. and S.K.J.; formal analysis, W.A.; funding acquisition, W.A., S.K.J. and N.A.; investigation, W.A.; methodology, W.A. and S.K.J.; project administration, S.K.J.; resources, W.A.; software, W.A.; supervision, S.K.J. and N.A.; validation, S.K.J. and N.A.; visualization, W.A.; writing—original draft, W.A.; writing—review and editing, S.K.J. and N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Deputyship for Research and Innovation, Ministry of Education in Saudi Arabia, through project number (IFPRC-054-612-2020) and King Abdulaziz University, DSR, Jeddah, Saudi Arabia.

Data Availability Statement: The dataset is not publicly available, due to privacy and institutional restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hu, M.; Li, H. Student engagement in online learning: A review. In Proceedings of the 2017 International Symposium on Educational Technology, ISET 2017, Hong Kong, China, 27–29 June 2017; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2017; pp. 39–43. [CrossRef]
2. Altuwairqi, K.; Jarraya, S.K.; Allinjawi, A.; Hammami, M. A new emotion–based affective model to detect student's engagement. *J. King Saud Univ.-Comput. Inf. Sci.* **2021**, *33*, 99–109. [CrossRef]
3. Zhang, Z.; Li, Z.; Liu, H.; Cao, T.; Liu, S. Data-driven Online Learning Engagement Detection via Facial Expression and Mouse Behavior Recognition Technology. *J. Educ. Comput. Res.* **2020**, *58*, 63–86. [CrossRef]

4. Mohamad Nezami, O.; Dras, M.; Hamey, L.; Richards, D.; Wan, S.; Paris, C. Automatic Recognition of Student Engagement Using Deep Learning and Facial Expression. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 273–289. [CrossRef]
5. Khenkar, S.; Jarraya, S.K. Engagement detection based on analyzing micro body gestures using 3D CNN. *Comput. Mater. Contin.* **2022**, *70*, 2655–2677. [CrossRef]
6. Kaur, A.; Mustafa, A.; Mehta, L.; Dhall, A. Prediction and Localization of Student Engagement in the Wild. In Proceedings of the 2018 Digital Image Computing: Techniques and Applications (DICTA), Canberra, ACT, Australia, 10–13 December 2018. Available online: http://arxiv.org/abs/1804.00858 (accessed on 26 June 2018).
7. Murshed, M.; Dewan, M.A.A.; Lin, F.; Wen, D. Engagement Detection in e-Learning Environments using Convolutional Neural Networks. In Proceedings of the 2019 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), Fukuoka, Japan, 5–8 August 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 80–86. [CrossRef]
8. Wang, Y.; Yao, Q.; Kwok, J.T.; Ni, L.M. Generalizing from a Few Examples: A Survey on Few-shot Learning. *ACM Comput. Surv.* **2020**, *53*, 63. [CrossRef]
9. Singh, M.; Hoque, X.; Zeng, D.; Wang, Y.; Ikeda, K.; Dhall, A. Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. In Proceedings of the 25th International Conference on Multimodal Interaction, Paris, France, 9–13 October 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 174–182. [CrossRef]
10. Liu, Y.; Zhang, H.; Zhang, W.; Lu, G.; Tian, Q.; Ling, N. Few-Shot Image Classification: Current Status and Research Trends. *Electronics* **2022**, *11*, 1752. [CrossRef]
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
12. Dan, J.; Liu, Y.; Xie, H.; Deng, J.; Xie, H.; Xie, X.; Sun, B. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023.
13. Mandia, S.; Singh, K.; Mitharwal, R. Vision Transformer for Automatic Student Engagement Estimation. In Proceedings of the 2022 IEEE 5th International Conference on Image Processing Applications and Systems (IPAS), Genova, Italy, 5–7 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
14. Hasnine, M.N.; Bui, H.T.T.; Tran, T.T.T.; Nguyen, H.T.; Akçapõnar, G.; Ueda, H. Students' emotion extraction and visualization for engagement detection in online learning. In *Procedia Computer Science*; Elsevier B.V.: Amsterdam, The Netherlands, 2021; pp. 3423–3431. [CrossRef]
15. Snell, J.; Swersky, K.; Zemel, T.R. Prototypical Networks for Few-shot Learning. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
16. Zhang, H.; Fu, Y.; Meng, J. Engagement Detection in Online Learning Based on Pre-trained Vision Transformer and Temporal Convolutional Network. In Proceedings of the 2024 36th Chinese Control and Decision Conference (CCDC), Xi'an, China, 25–27 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1310–1317. [CrossRef]
17. Maddula, N.V.S.S.; Nair, L.R.; Addepalli, H.; Palaniswamy, S. Emotion Recognition from Facial Expressions Using Siamese Network. In *Communications in Computer and Information Science*; Springer Science and Business Media Deutschland GmbH: Berlin/Heidelberg, Germany, 2021; pp. 63–72. [CrossRef]
18. Liu, B.; Yu, X.; Yu, A.; Zhang, P.; Wan, G.; Wang, R. Deep Few-Shot Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2290–2304. [CrossRef]
19. Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to Compare: Relation Network for Few-Shot Learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1199–1208.
20. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021.
21. Su, R.; He, L.; Luo, M. Leveraging part-and-sensitive attention network and transformer for learner engagement detection. *Alex. Eng. J.* **2024**, *107*, 198–204. [CrossRef]
22. Gretton, A.; Borgwardt, K.M.; Rasch, M.; Schölkopf, B.; Smola, A.J. A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems*; Schölkopf, B., Platt, J., Hoffman, T., Eds.; MIT Press: Cambridge, MA, USA, 2006; Volume 19.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021.
24. Selim, T.; Elkabani, I.; Abdou, M.A. Students Engagement Level Detection in Online e-Learning Using Hybrid EfficientNetB7 Together With TCN, LSTM, and Bi-LSTM. *IEEE Access* **2022**, *10*, 99573–99583. [CrossRef]

25. Mandia, S.; Singh, K.; Mitharwal, R.; Mushtaq, F.; Janu, D. Transformer-Driven Modeling of Variable Frequency Features for Classifying Student Engagement in Online Learning. *arXiv* **2025**, arXiv:2502.10813.
26. Tieu, B.H.; Nguyen, T.T.; Nguyen, T.T. Detecting Student Engagement in Classrooms for Intelligent Tutoring Systems. In Proceedings of the 2019 6th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam, 12–13 December 2019; pp. 145–149. [CrossRef]
27. Abedi, A.; Khan, S.S. Detecting Disengagement in Virtual Learning as an Anomaly using Temporal Convolutional Network Autoencoder. *Signal Image Video Process.* **2023**, *17*, 3535–3543. [CrossRef]