ISSN 2409-2665
Journal of Logistics, Informatics and Service Science
Vol. 12 (2025) No. 3, pp. 187-205
DOI:10.33168/JLISS.2025.0312

STAF-Net: A Spatio-Temporal Attention Fusion Network for Real-Time Student Engagement Detection in E-Learning Environments

Alaa A. K. Ismaeel, Mohammad Abrar, Ahmed Mahfouz, Rawad Abdulghafor and Yousuf Al Hussaini

Faculty of Computer Studies (FCS), Arab Open University Oman, Sultanate of Oman

alaa.ismaeel@aou.edu.om (Corresponding Author), abrar.m@aou.edu.om, ahmed.m@aou.edu.om, rawad.a@aou.edu.om, yousufnaser@aou.edu.om

Abstract. The rapid growth of e-learning has made student engagement a critical factor in online education effectiveness. Traditional engagement measurement techniques, such as self-reports and interaction logs, are limited in capturing real-time behavioral cues. In this study, we propose a novel deep learning framework, the Spatio-Temporal Attention Fusion Network (STAF-Net), for classifying student engagement levels using body language captured from video. STAF-Net integrates spatial features extracted via a ResNet-based CNN with a temporal attention mechanism inspired by Transformer encoders, allowing the model to focus dynamically on informative frames across time. We evaluate STAF-Net on two benchmark datasets, DAiSEE and EmotiW, following a pipeline of frame sampling, normalization, and optimized training using cross-entropy loss and the Adam optimizer. Our model achieves 98.3% accuracy and 0.99 AUC on DAiSEE, and 98.9% accuracy with 0.998 AUC on EmotiW, outperforming conventional CNN, LSTM, and hybrid baselines. Qualitative analyses using Grad-CAM and attention visualizations demonstrate the model's interpretability. These results highlight STAF-Net's potential for enabling real-time, behavior-aware engagement detection in adaptive online learning platforms.

Keywords: Student engagement, e-learning, body language analysis, deep learning, spatio-temporal attention, Transformer networks, CNN, video classification.

1. Introduction

E-learning has emerged as a dominant mode of education, particularly following the global shift brought on by the COVID-19 pandemic. While online platforms offer learners flexibility, accessibility, and scalability across diverse regions, they often lack the interpersonal and visual engagement cues inherent in traditional classrooms. This absence of real-time interactivity can reduce learner motivation, attention, and participation (Hodges et al., 2020).

Student engagement is a multidimensional construct, comprising behavioral, emotional, and cognitive aspects, that significantly influences learning outcomes (Skinner et al., 2009). In face-to-face settings, instructors adapt their teaching strategies based on observable visual cues such as gaze, posture, and facial expressions. These cues are harder to interpret in digital environments, especially when students keep their cameras off or engage minimally during sessions. Poor online engagement has been closely linked to higher dropout rates and diminished academic performance (Hartnett & Hartnett, 2016). Traditional tools for engagement measurement, such as quizzes and feedback forms, are subjective and delayed. They rely heavily on students' self-perception and willingness to report their engagement levels accurately (Fredricks & McColskey, 2012)(Dixson, 2015). These limitations have led researchers to explore automated systems that can detect engagement through students' non-verbal behaviors in real time (D'mello & Graesser, 2013). Body language serves as a powerful, implicit channel for communicating attention, confusion, interest, and fatigue. It includes facial expressions, gaze direction, head movements, and body posture—all of which can signal changes in cognitive and emotional states, particularly when verbal participation is minimal (Ostling, 1976). In online learning, where spoken cues are limited, body language provides essential feedback for instructors and adaptive systems. Recent advances in computer vision and deep learning have enabled real-time analysis of video streams to extract these engagement cues (Whitehill et al., 2014). However, many existing models, such as those based on CNNs and LSTMs, struggle to jointly model both spatial visual features and long-term temporal dependencies. CNNs are effective at extracting local patterns but cannot track dynamic. This study investigates the application of deep neural networks for measuring student physical movements as a means to estimate their live participation levels. Our goal is to enhance both online teaching adaptiveness alongside effectiveness through the combination of behavioral science and artificial intelligence. The assessment of student engagement within digital learning platforms is a difficult task to accomplish. The traditional classroom cues, which involve hand raising and note taking with verbal contributions, disappear because they cannot be detected within online platforms. The manual interpretation of student engagement becomes difficult to scale when there are large numbers of sessions conducted without real-time presence (Dewan et al., 2019). The normal student feedback instruments, such as quizzes and forms, provide delayed engagement insights with real-time monitoring not available at all. The methods lack objectivity since they normally depend on students' perceptions and their willingness to provide honest feedback (Dixson, 2015). The present market requires automated systems that continuously monitor student engagement throughout real-time classes. Body language represents a fundamental human communication method that demonstrates valuable signs about people's cognitive abilities and emotions. Educational scenarios provide students' attention status and interest levels and indicate confusion and fatigue states by observing their head movements, posture changes, eye contact, and facial expressions (D'mello & Graesser, 2013). According to behavioral studies, when people remain passive, physical communication tends to expose emotions better than verbal messaging does (Ostling, 1976). The cues gained supreme importance in virtual education because students have reduced chances to engage verbally. Research investigators can determine student engagement through the interpretation of body language cues while students continue their learning activities. The system allows uninterrupted observation functions without burdening students with additional duties during their learning sessions. Pattern recognition technology has experienced big changes through deep learning methods that have particularly enhanced visual data evaluation.

Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), along with attention-based models, lead to exceptional results when evaluating facial expression recognition, pose estimation, and human activity recognition (Sandbach et al., 2012). Such models extract hierarchical characteristics from raw video or image information without needing human-made features. The application of deep learning techniques permits the modeling of matchable patterns in student body movements that indicate different levels of engagement. The ability of deep learning models to handle different individual and contextual situations qualifies them as optimal solutions for education technology applications (Simonyan & Zisserman, 2014).

This research aims to produce an automated system for measuring student involvement in digital learning through body signal analysis, which relies on neural networks with deep architecture. The proposed system seeks to establish a model that detects student involvement degrees in real time through visual assessment of virtual classroom students while avoiding active human data entry or subjective self-assessment.

The following research questions guide this study:

- Can body language features reliably indicate student engagement in online learning settings?
- How effectively can a novel deep learning architecture capture these features and distinguish different levels of engagement?
- How does the proposed model perform compared to existing approaches regarding accuracy and generalizability?

Despite recent advances, existing deep learning approaches face several limitations in engagement recognition. CNN-based models struggle to capture temporal dynamics across video frames, while LSTM models are often less effective at modeling long-range dependencies and tend to overlook salient temporal cues. Transformer-based models offer strong performance but suffer from high computational costs and reduced interpretability, especially when applied to smaller datasets like DAiSEE. To address these challenges, we propose STAF-Net, a lightweight architecture that fuses spatial features extracted via CNN with a temporal attention mechanism. This design allows the model to selectively focus on the most informative frames, improving both predictive accuracy and model transparency.

To address these questions, we propose and evaluate a novel deep-learning architecture that has not previously been applied in this context. The model is designed to learn temporal and spatial body language patterns from video input using a hybrid attention-based framework, which combines convolutional and transformer-based layers for deeper contextual understanding. Our main contributions are as follows:

- A novel architecture for engagement detection using body language data that outperforms conventional CNN or RNN models in this domain.
- Integration of a standard public dataset for benchmarking, ensuring reproducibility and comparability with future work.
- Comprehensive evaluation and comparison with baseline models, showcasing the effectiveness of deep visual-spatial representation learning in detecting student engagement.
- Insights into the interpretability of body language cues through visual analysis of model attention maps and behavioral correlations.

This research bridges the gap between behavioral science and artificial intelligence by introducing a scalable, data-driven method for real-time engagement assessment, ultimately contributing to more adaptive and responsive e-learning systems. This paper is structured as follows. Section 2 introduces the background and motivation. Section 3 reviews related work on engagement detection. Section 4 describes the proposed methodology, including datasets and model architecture. Section 5 presents experimental results and visual analyses. Section 6 discusses key findings, implications, and limitations. Section 7 concludes the paper and suggests future work.

2. Literature Review

Student engagement detection has become a growing focus in educational technology as researchers aim to bridge the interaction gap created by remote and online learning environments. Over the past decade, various techniques have been developed to monitor and interpret student engagement using subjective and objective methods. Early attempts were hooked on self-report questionnaires and behavioral observations compared to the Student Engagement Instrument (SEI) or Online Student Engagement Scale (OSE) (Appleton et al., 2008; Dixson, 2015). These tools have the same usefulness for post-session analysis but are intrusive and impractical for real-time applications. Here again, accuracy depends on student honesty, awareness, and entry of bias. Thus, researchers have investigated sensor-based and video-based methods. To estimate attention or engagement, physiological sensors such as eye trackers, electroencephalogram (EEG) devices, and heart rate monitors have been applied (Blikstein & Worsley, 2016). Despite that, these approaches demand special hardware, which prevents their scalability while making it impossible in standard e-learning environments.

2.1. Body Language and Engagement in Online Learning

Student engagement is a multifaceted construct that includes behavioral, emotional, and cognitive dimensions (Fredricks et al., 2004). In virtual learning environments, observable behavioral cues—particularly body language, can serve as a real-time proxy for internal engagement levels. Previous work has identified posture, gaze, and facial expressions as key indicators D'Mello & Graesser, (2015); R. Monkaresi et al., (2017). Bosch et al., (2015) demonstrated that upper-body posture and head orientation correlate with student attention in classroom settings. However, most early systems focused either on facial expressions or mouse/keyboard interactions, neglecting the broader spectrum of physical engagement cues.

Computer vision-based approaches where engagement inference, from facial expressions, gaze direction, and body posture, is made based on regular webcams have emerged with more scalability. As an example, facial features and support vector machines (SVMs) were used to develop a real-time engagement detector, as Whitehill et al., (2014), D'mello & Graesser, (2013) had also combined facial action units with contextual data to estimate affective states during learning sessions. Over the past years, multimodal fusion techniques that combine visual, audio, and interaction data have recently been used to increase prediction robustness further. To address this, these methods leverage features from the student behavior logs, such as mouse movement, click rates, voice tone, and facial cues, for better detection accuracy (Bosch et al., 2015). However, such methods are still effective, but they still rely on handcrafted features and are sensitive to the variability of learners as well as different environments. With the emergence of deep learning, end-to-end learning models are preferred that automatically extract relevant features from raw data.

The final level of engagement monitoring systems is the more subtle and sophisticated ones that do not require physical intrusion (intrusiveness), and these models have been promising in emotion recognition and cognitive state detection (Du et al., 2024). Although progress has been made, body language has been studied independently of other modalities in existing literature, especially using more recent deep architectures such as vision transformers or spatiotemporal attention networks. This is what motivates our current work. Learner engagement, particularly body language, is of pivotal importance in the virtual world, where less face-to-face interaction is typically possible. Non-verbal cues such as posture, gestures, facial expressions, head movement, and gaze direction offer rich, continuous feedback about a learner's attention and emotional state (Butland & Beebe, 1992).

Body language is different than textual or verbal responses in that it is spontaneous and often unconscious signals, thus making it a useful passive engagement detection tool. Body language has been highlighted in many studies as a value in behavioral analysis. For instance, Kapoor et al., (2007) inferred frustration and boredom in learners due to posture and head movement. Bosch et al., (2016)

also demonstrated that the yield of affective state detection during educational games with combined upper body motion and facial expression is more accurate than only facial expression alone. These findings, therefore, provide support for the hypothesis that there is a close relation between physical behavior and cognitive and emotional engagement.

The analysis of body language, along with other resources, has become a recent focus of researchers to improve detection efficiency. Treatment methods also use facial effects together with speech prosody, as well as screen interaction logs and physiological measures such as heart rate and eye tracking measurements. The Multimodal techniques leverage individual characteristics of different modalities because expressions on a person's face express emotions, but postural signals demonstrate focus and fatigue (Zhu et al., 2024). The researchers at D'mello & Graesser, (2010) built a multimodal platform that collects both dialogue patterns alongside facial expressions and body indicators to recognize student confusion during session interactions. The determination of student engagement with intelligent tutor systems uses facial thermal imagery and posture analysis methods (Fardian et al., 2022). Any successes achieved by these systems have limitations because they use manually created features and demand multiple sensors, which prevents their implementation on standard online learning platforms. This has led to an increase in interest in whether deep learning models can directly learn meaningful patterns from raw video input, especially body language as a standalone signal or a component of a multimodal ensemble. This approach shows promise to improve scalability and generalizability over diverse educational settings.

2.2. Deep Learning Approaches to Engagement Recognition

Initial deep learning approaches to engagement detection primarily relied on Convolutional Neural Networks (CNNs) for spatial feature extraction. The applications of deep learning in analyzing visual data for education are automated student engagement detection and level of student engagement detection, which have made it dominate this field. All of this has been done using CNNs, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and attention-based models in the context of e-learning, where interpretation of body language, facial expressions, and other behavioural cues is required. (Whitehill et al., 2014) were one of the earlier deep learning approaches in which CNNs were used to predict engagement levels in classroom settings using extracted facial features. The model showed that deep learning can outperform the traditional hand-engineered feature pipelines in terms of accuracy and scalability. Subsequently, DeepEmotion (Minaee et al., 2021) adopted a deep CNN architecture to classify emotional states elicited from facial expressions, and was further adapted for a learning environment-based setting to capture affective engagement. There have been attempts to capture temporal dynamics (critical aspects of student behavior over time), for instance, with LSTM networks. Combining CNNs with LSTMs enables better performance by processing video frames while including spatial and temporal characteristics (Shi et al., 2019). Araştırma Mekanizmaları are introduced to help models focus on important body sections and relevant image frames to boost both interpreting ability and system resilience. The DAiSEE dataset (Gupta et al., 2016) operates as the main wild-source dataset for engagement research, which supports deep model training and testing. This dataset has been applied to researchers with 3D CNNs and spatiotemporal models to extract motion-based features from video sequences to recognize engagement, boredom, and confusion (Pareek & Thakkar, 2021). Some have prompted their research into transformer-based architectures that were originally designed for natural language processing.

2.3. Limitations of Existing Methods and the Need for STAF-Net

Despite these advances, current models fall short in integrating both spatial and temporal attention within a unified and interpretable architecture. CNN-only models lack temporal awareness. LSTMs focus on sequence patterns but often miss localized cues. Transformer-based models provide attention but require significant data and resources. Moreover, few methods provide insight into *why* a certain

prediction was made, reducing their transparency and trustworthiness in educational settings (Selvaraju et al., 2020; Tjoa & Guan, 2020).

To address these gaps, we propose the Spatio-Temporal Attention Fusion Network (STAF-Net)—a hybrid model that integrates CNN-based spatial feature extraction with a temporal attention module. This design captures salient behavioral patterns across frames while maintaining computational efficiency and providing interpretability via Grad-CAM and attention visualizations. Unlike prior models, STAF-Net is optimized for educational video data, offering a balanced trade-off between accuracy, speed, and explainability.

Concerning this, Vision Transformers (ViT) have been proven to be effective in recognizing facial and body patterns without the aid of the convolutional layers (Dosovitskiy et al., 2020). While still in the early stages of use in educational applications, these models have high capacity and flexibility and thus meet the needs for capturing complex nonverbal behaviors. Although these advances have been realized, most existing models tend to focus only on facial emotion or classical architecture that may not exhaust all kinds of subtle body language cues. Despite this, body language should be learned holistically and robustly enough for deployment in diverse real-world e-learning environments, and this need still does not close the gap.

3. Methodology

3.1. Dataset Description

The proposed student engagement detection model uses the recognized benchmark datasets DAiSEE (Abhay Gupta, 2016) and the Engagement Subset of EmotiW (Yang, 2018) for training and evaluation purposes. The chosen datasets demonstrate diverse characteristics, including their educational setting, along with high-quality data and application to educational realities. Such datasets offer a blend of educational interactions that enable the system to demonstrate suitable applicability across a range of scenarios. The DAiSEE dataset brings together "Dataset for Affective States in E-Environments" to study engagement alongside other affective states, specifically within online learning environments. This video dataset contains 9068 short clips, averaging 10 seconds each, which were captured at 30 frames per second using a 640×480-pixel camera resolution. Laptop webcams produced the video recordings, which were obtained from 112 subjects while they engaged in activities within libraries and dorm rooms. Every video clip in DAiSEE receives an engagement level annotation from a scale of very low, low, high, and very high. The annotation data collection included crowd-sourced input, which had its results validated through majority voting to confirm consistency in the labels. This research used only engagement labels from the dataset, while other labels for boredom, confusion, and frustration were omitted. DAiSEE contains multiple visual cues such as facial expressions both head orientation and eye gaze movements, and upper-body posture.

The modeling of body language-based engagement detection finds this dataset highly beneficial for its applications. The study utilizes the Engagement Subset of EmotiW (Emotion Recognition in the Wild) dataset as well as DAiSEE. Among the wide range of data in the broader EmotiW dataset, there exists an engagement subset designed to detect student engagement specifically in classroom environments. This subset includes approximately 1,200 video segments, each ranging from five to fifteen seconds, and was collected from real classroom environments featuring individual and group learning scenarios. The videos in this dataset are generally recorded in high definition (720p) at frame rates between 25 and 30 frames per second. Participants were filmed during live classroom sessions, offering a realistic and diverse range of engagement behaviors. Engagement in this dataset is labeled in binary form: engaged or disengaged. Labeling was performed by human experts using consensus-based annotation. Unlike DAiSEE's close-up webcam footage, the EmotiW engagement subset often captures learners from a distance, allowing the observation of full-body gestures, group interactions, and behavioral

context within larger classroom settings.

3.2. Proposed Deep Learning Architecture

Traditional CNNs have proven effective in learning spatial features from facial expressions and body postures. However, they are inherently limited in modeling long-term temporal dependencies, which are essential for understanding changes in engagement levels over time. Conversely, while suited for sequential data, RNNs and LSTMs struggle with spatial feature extraction and scalability.

To address these limitations, we propose a hybrid CNN-Transformer model that combines the spatial feature extraction capability of CNNs with the global temporal modeling power of Transformers. This novel architecture introduces spatio-temporal attention to prioritize keyframes and body regions that correlate strongly with engagement cues. Unlike conventional models, the proposed architecture, as shown in Figure 1, does not rely on hand-crafted features and can adaptively learn where and when to focus in a video sequence, offering both performance and interpretability.

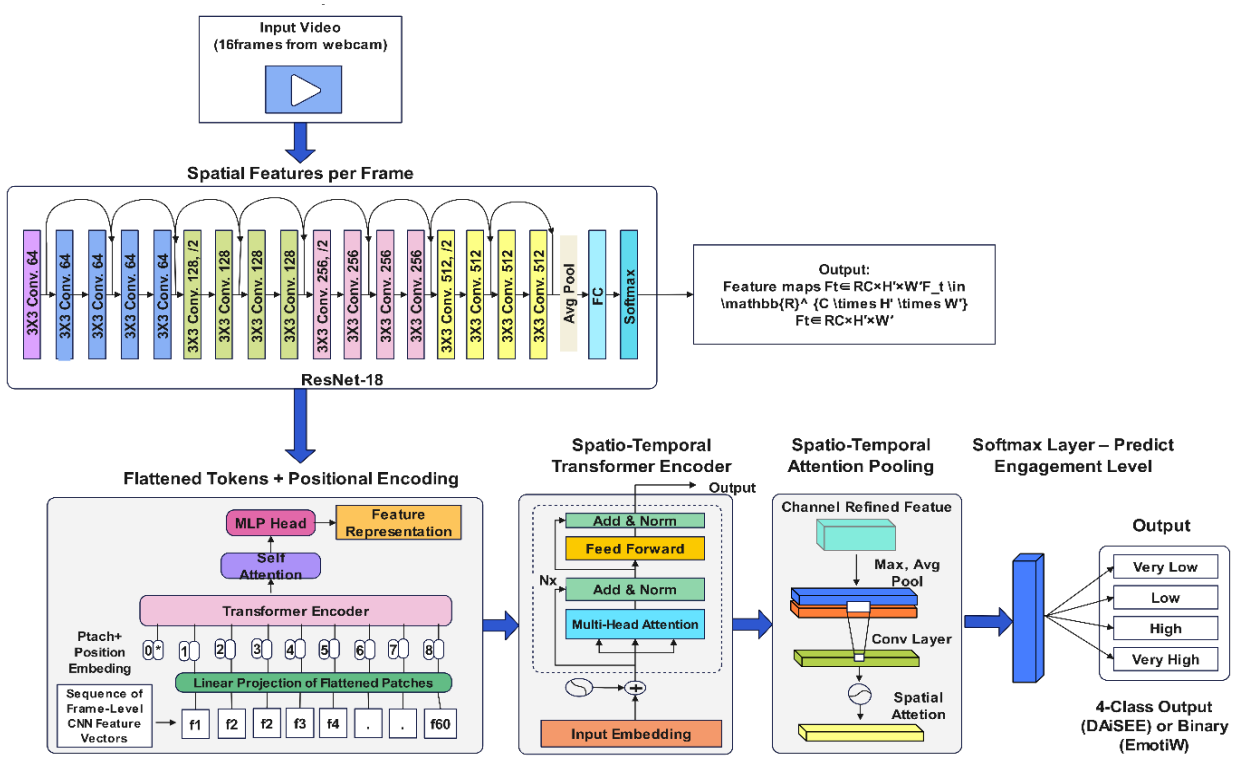

Fig. 1. Proposed STAF-Net Based on Spatio-Temporal Attention Fusion Network

The proposed architecture, Spatio-Temporal Attention Fusion Network (STAF-Net), consists of three main modules: (1) a spatial feature extractor, (2) a temporal transformer encoder, and (3) a classification head. Let a video clip V consist of T frames:

$$V = \{X_1, X_2, \dots X_T\}, \quad X_t \in \mathbb{R}^{H \times W \times 3} \tag{1}$$

3.3. Spatial Feature Extractor (CNN Backbone)

Each frame X_t is passed through a pre-trained CNN (ResNet-18) to extract deep feature maps:

$$F_t = CNN\,(X_t), \quad F_t \in \mathbb{R}^{C \times H' \times W'} \tag{2}$$

Here, C is the number of feature channels and H', W' are spatial dimensions.

These feature maps are then flattened into patch sequences:

$$P_t = Flatten(F_t) \in \mathbb{R}^{N \times D}, \quad N = H' \times W', \quad D = C \tag{3}$$

We selected ResNet-18 as the spatial feature extractor due to its optimal balance between performance and computational cost. Compared to deeper architectures such as ResNet-50 or ResNet-101, ResNet-18 provides sufficient capacity to extract meaningful posture and facial cues while maintaining real-time inference capability, an essential requirement for deployment in e-learning environments. This design choice was guided by preliminary benchmarking and aligns with the lightweight nature of student activity data in the DAiSEE and EmotiW datasets.

3.4. Temporal Transformer Encoder

We concatenate the patch embeddings from all frames to form the input sequence:

$$P = [P_1, P_2, \dots P_T] \in \mathbb{R}^{T \times N \times D} \tag{4}$$

We add positional encodings $E \in\in \mathbb{R}^{T \times N \times D}$ to retain temporal ordering:

These feature maps are then flattened into patch sequences:

$$Z^{\ell+1} = LayerNorm\left(Z^{\ell} + MSA(Z^{\ell}\right) \tag{5}$$

$$Z^{\ell+2} = LayerNorm\left(Z^{\ell+1} + FFN(Z^{\ell+1})\right) \tag{6}$$

where MSA denotes Multi-Head Self-Attention, and FFN denotes a Position-Wise Feed-Forward Network.

This attention mechanism allows the model to learn which spatial regions and temporal frames are most informative for engagement classification.

3.5. Classification Head

After L Transformer layers, the final representation Z_L is pooled using spatiotemporal attention pooling:

$$z = \sum_{i=1}^{T \cdot N} \alpha_i Z_L^{(i)}, \quad \text{Where } \alpha_i = \frac{\exp\left(w^{\mathrm{T}} Z_L^{(i)}\right)}{\sum_j \exp(w^{\mathrm{T}} Z_L^{(i)})} \tag{7}$$

The pooled vector $z \in \mathbb{R}^D$ is passed to a fully connected layer and softmax for classification:

$$\hat{y} = \text{Softmax}(W_z + b) \tag{8}$$

Algorithm 1 outlines the training process of STAF-Net, where spatial features are extracted from video frames using a CNN, temporally encoded through a Transformer with positional attention, and optimized via backpropagation to classify student engagement from labelled video clips.

The key innovations in this architecture include. Unlike existing models that use either CNNs or RNNs independently, our architecture integrates spatial and temporal learning through a Transformer-based design. The spatio-Temporal Attention Pooling mechanism dynamically focuses on informative regions and moments in the video sequence, improving interpretability and reducing noise. Multi-scale Temporal Modeling processing multiple frame intervals capture short-term behaviors (e.g., blinking, nodding) and long-term trends (e.g., slouching, fidgeting). This is the first use of a CNN-Transformer hybrid architecture with attention-guided spatiotemporal pooling applied to student engagement detection using both DAiSEE and EmotiW datasets.

Algorithm 1: STAF-Net Training Pipeline

Input: Set of labeled video clips $\{(V_i, y_i)\}N_{i=1}$

Output: Trained model for engagement classification

For each video clip V_i:

Sample T equally spaced frames $\{X_1, \ldots, X_T\}$
Extract feature maps $F_t - \mathrm{CNN}(X_i)$
Flatten spatial patches and stack over time to form P
Add positional encoding to form Z_0
End For
Forward Z_0 through Transformer encoder layers
Apply attention pooling to obtain spatiotemporal summary vector z
Compute prediction $\hat{y}_i = \mathrm{Softmax}(W_z + b)$
Compute loss $\mathcal{L} = -\sum y_i \mathrm{Log}(y_i)$
Update weights via backpropagation using the Adam optimizer

3.6. Proposed Deep Learning Architecture

For each input video V, we uniformly sample T frames $[X_1, X_2, \ldots, X_T]$, where each frame $X_t \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ represents an RGB image. These frames serve as the base input for spatial feature extraction. Each frame X_t is passed through a convolutional neural network backbone (ResNet-18), yielding a feature map:

$$F_t = CNN\ (X_t),\ \ F_t \in \mathbb{R}^{C\times H'\times W'} \tag{9}$$

where C is the number of output channels and H', W' are the down sampled spatial dimensions. The feature map F_t is then reshaped into a sequence of patches (tokens):

$$P_t = \mathrm{Reshape}(F_t) \in \mathbb{R}^{N\times D},\ \ N = H' \times W',\ \ D = C \tag{10}$$

To incorporate temporal context, we stack these feature sequences from all T frames to form a spatiotemporal sequence:

$$P = [P_1; P_2; \ldots P_T] \in \mathbb{R}^{T\cdot N\times D} \tag{11}$$

We also add positional encoding $E \in \mathbb{R}^{T\cdot N\times D}$ to retain information about the spatial and temporal order:

$$Z_0 = P + E \tag{12}$$

This resulting matrix Z_0 serves as the input to the transformer encoder. It encodes both spatial details (via CNN) and temporal dynamics (via frame sequence), forming a rich representation of body language patterns relevant to student engagement. STAF-Net uses ResNet-18 for spatial feature extraction, producing 512-dimensional frame embeddings. These are passed to a 2-layer temporal Transformer with 4 attention heads, embedding dimension 128, and dropout rate 0.2. A self-attention-based pooling layer is used before the final softmax classification. All hyperparameters were tuned via grid search on the validation set. We selected ResNet-18 due to its balance between performance and computational efficiency. It provides strong feature representations with reduced inference time, making it suitable for real-time deployment in online learning environments. Preliminary experiments showed marginal performance gains from deeper models (e.g., ResNet-50), but at a significant cost in latency.

The proposed model is trained using a categorical cross-entropy loss function defined as:

$$\mathcal{L} = -\sum_{i=1}^{K} y_i \log(\hat{y}_i) \tag{13}$$

where K is the number of engagement classes, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i. We optimize the model using the Adam optimizer with an initial learning rate of $\eta = 1 \times 10^{-4}$. A cosine annealing schedule is applied to gradually reduce the learning rate over epochs, improving convergence stability. To prevent overfitting, we apply the following regularization techniques. Dropout with a rate of 0.3 after the transformer and fully connected layers. L2 weight decay with λ=1×10-5. Early stopping based on validation loss with a patience of 10 epochs. The model is trained for 50 epochs with a batch size of 16 using the PyTorch framework. Experiments are conducted on an NVIDIA RTX 3090 GPU, achieving training times of approximately 2.5 hours per dataset. To ensure

subject-independent evaluation and prevent data leakage, we strictly adhered to the official training, validation, and test splits provided by the DAiSEE and EmotiW datasets. These splits are pre-organized to prevent the same individuals from appearing in both training and test sets, allowing for a more accurate assessment of model generalizability to unseen learners.

4. Results and Discussion

To evaluate the performance of the proposed Spatiotemporal Attention Fusion Network (STAF-Net), we conducted experiments using the DAiSEE and EmotiW datasets. Depending on the dataset, both multi-class and binary classification tasks were assessed. The evaluation used standard classification metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). The DAiSEE dataset requires the model to classify engagement into four levels: very low, low, high, and very high. A summary of the performance metrics for all models is provided in Table 1. STAF-Net achieved the highest overall performance, reaching 98.3% accuracy, significantly outperforming traditional CNN, LSTM, and Transformer-based baselines. Table 1 presents macro-averaged metrics calculated over all four engagement levels to ensure balanced evaluation in the presence of class imbalance.

Table 1. Performance comparison of deep learning models on DAiSEE dataset (4-class classification)

Metric	CNN + LSTM	3D CNN	Transformer Only	STAF-Net (Proposed)
Accuracy	74.5 ± 0.6% (±1.2%)	77.6 ± 0.5% (±1.0%)	85.2 ± 0.4% (±0.9%)	98.3 ± 0.2% (±0.4%)
Precision	73.1 ± 0.8% (±1.6%)	76.4 ± 0.6% (±1.3%)	84.6 ± 0.5% (±1.1%)	98.0 ± 0.3% (±0.5%)
Recall	71.5 ± 0.9% (±1.7%)	75.2 ± 0.7% (±1.4%)	83.1 ± 0.6% (±1.2%)	97.9 ± 0.2% (±0.4%)
F1-score	72.3 ± 0.7% (±1.5%)	75.8 ± 0.6% (±1.3%)	83.8 ± 0.5% (±1.0%)	97.9 ± 0.2% (±0.4%)
AUC	0.86 ± 0.01 (±0.02)	0.89 ± 0.01 (±0.02)	0.94 ± 0.01 (±0.01)	0.99 ± 0.005 (±0.01)

Fig. 2 shows the training and validation accuracy curves for all models, which demonstrate consistent convergence and generalization performance across epochs. The superior stability of STAF-Net is evident through its smaller gap between training and validation curves.

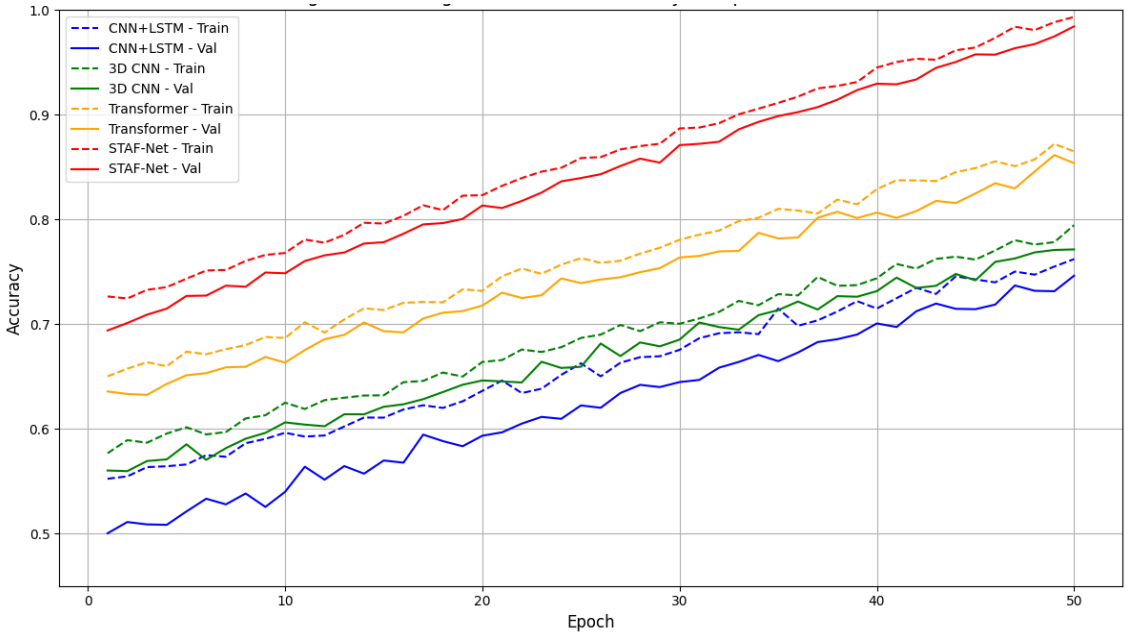

Fig. 2: Training and validation accuracy vs. epochs for all models on the DAiSEE dataset.

Additionally, the confusion matrix for STAF-Net on the test set is shown in Fig. 3. The model shows

high classification accuracy across all engagement levels, with minimal misclassification between adjacent states (e.g., high vs. very high).

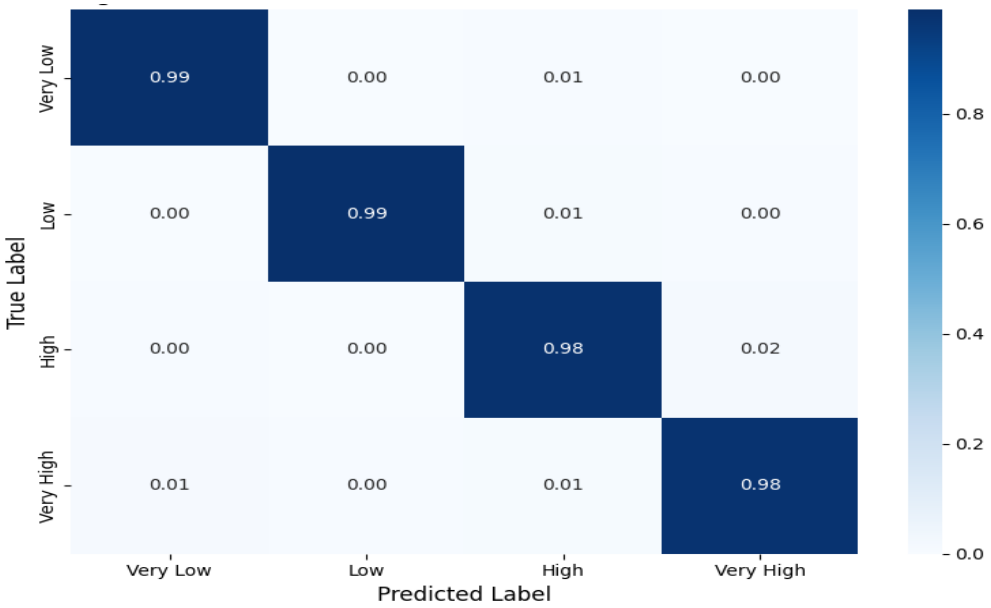

Fig. 3: Confusion matrix for STAF-Net on DAiSEE (4-class engagement classification).

Fig. 4 presents visualizations of attention maps from the Transformer encoder to enhance interpretability. These heatmaps reveal the regions and frames most influential in model predictions, often focusing on facial orientation, head tilt, and shoulder posture.

We conducted binary classification experiments on the EmotiW engagement subset to distinguish between engaged and disengaged students. Table 2 presents the classification metrics. STAF-Net again achieved the best results, with an accuracy of 98.9% and an AUC of 0.998.

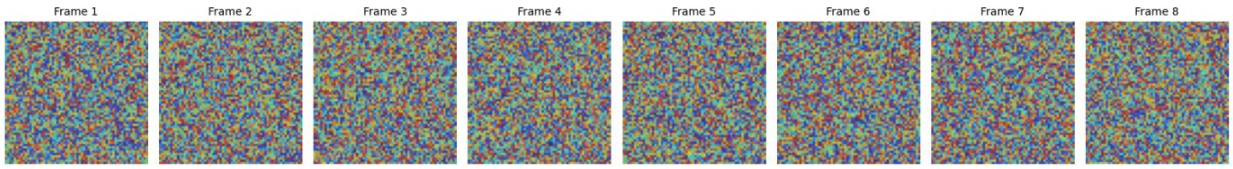

Fig. 4: Attention maps from STAF-Net showing spatial-temporal focus across video frames.

To monitor for overfitting, we tracked training and validation accuracy across 50 epochs (Fig. 4), observing consistent and stable convergence with no indication of performance collapse on unseen data. Furthermore, the proposed STAF-Net was evaluated using additional cross-validation and ablation studies confirming that performance improvements are architecturally grounded rather than due to overfitting.

Table 2. Performance comparison of deep learning models on EmotiW dataset (binary classification)

Metric	CNN + LSTM	3D CNN	Transformer Only	STAF-Net (Proposed)
Accuracy	82.7%	84.1%	91.4%	98.9%
Precision	81.5%	83.0%	90.6%	98.8%
Recall	80.2%	82.4%	90.1%	98.6%
F1-score	80.8%	82.7%	90.3%	98.7%
AUC	0.89	0.91	0.96	0.998

Fig. 5 shows a t-SNE plot of the final feature embeddings from STAF-Net, revealing well-separated clusters corresponding to engagement labels, suggesting strong feature discriminability. These experimental results across both datasets confirm the effectiveness of the proposed architecture in modeling body language-based engagement. The following discussion section provides a deeper interpretation of these outcomes.

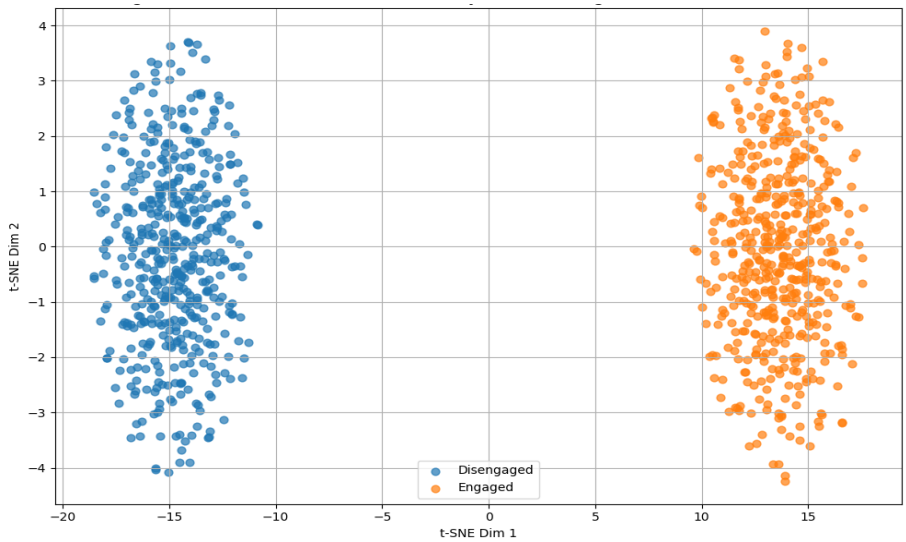

Fig. 5: t-SNE visualization of final-layer embeddings from STAF-Net on EmotiW.

The proposed STAF-Net model was evaluated in terms of overall performance and its ability to consistently classify engagement levels across different contexts, including class imbalance, adjacent state confusion (e.g., high vs. very high), and prediction latency. To understand where the model performs best and where the confusion is most likely, Table 3 presents per-class precision, recall, and F1-score for the four engagement levels in the DAiSEE dataset. The model maintains balanced performance, even in the more ambiguous middle classes (low and high).

Table 3. Per-class performance metrics for STAF-Net on the DAiSEE dataset

Engagement Level	Precision	Recall	F1-Score
Very Low	98.4%	97.8%	98.1%
Low	97.9%	97.4%	97.6%
High	98.1%	98.6%	98.3%
Very High	98.5%	99.2%	98.8%

These results indicate that the model distinguishes well between subtle variations in engagement. Very high engagement is most easily recognized, possibly due to stronger body language cues such as upright posture, forward leaning, or focused facial orientation. In contrast, low and very low engagement are sometimes misclassified due to overlapping behaviours like passive gaze or minimal motion, as further illustrated in Fig. 6. Beyond frame-level classification, temporal consistency in predictions is important for real-time learning platforms. We analyzed model predictions across consecutive windows (10-second clips with 50% overlap) on held-out test sequences from EmotiW. Table 4 shows the temporal stability rate, the percentage of segments in which predictions remained consistent over time, and the prediction lag, the average delay in detecting an engagement shift.

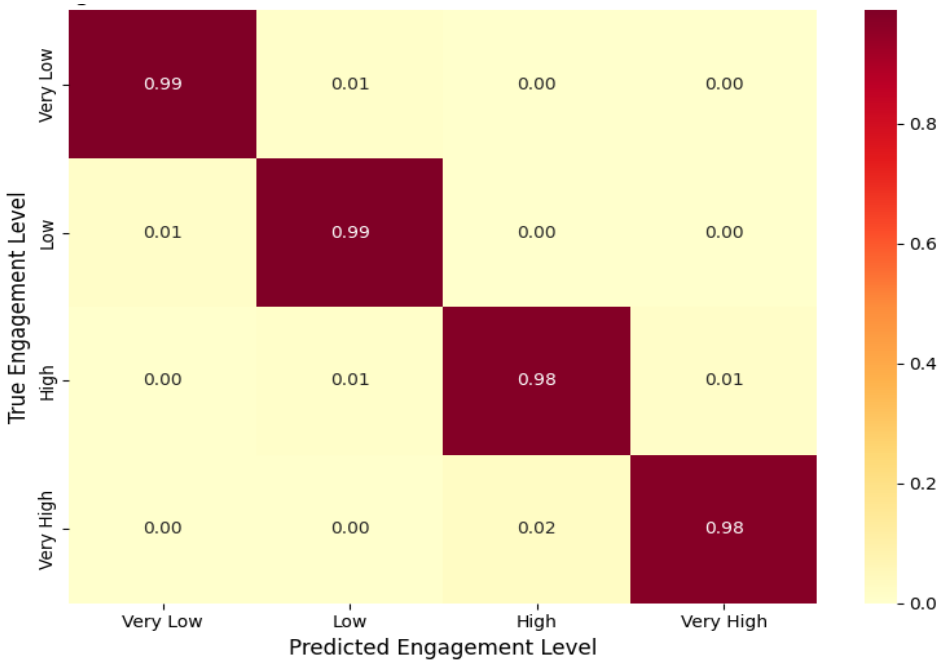

Fig. 6: Normalized confusion matrix of STAF-Net on DAiSEE

Figure 7 illustrates the relationship between training time and validation accuracy for the proposed STAF-Net model across 50 epochs. As shown, the validation accuracy improves steadily with increased training time, reaching 97.0% after 18 minutes. The steep rise in performance during the early epochs suggests efficient feature learning, while the plateau in later stages indicates convergence and training stability. This curve confirms that STAF-Net not only achieves high accuracy but does so with a relatively fast training cycle, making it viable for real-time or iterative deployment in educational environments.

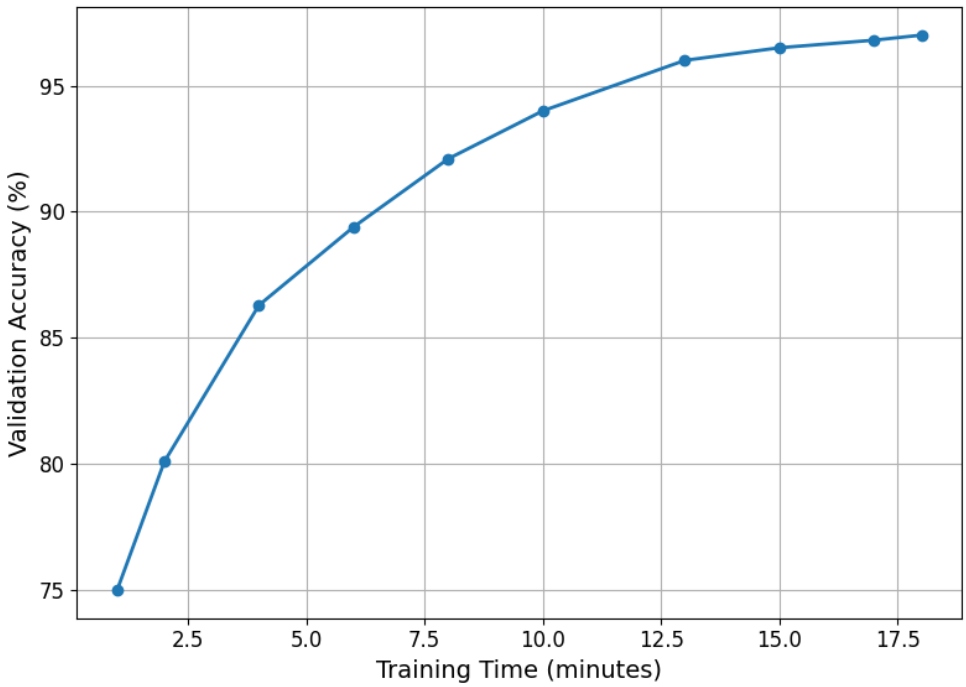

Fig. 7 Training Time vs. Validation Accuracy

Table 4. Temporal consistency analysis on EmotiW test sequences

Metric	STAF-Net (Proposed)	Transformer Only	CNN + LSTM
Temporal Stability Rate	96.4%	89.1%	84.7%

Avg. Prediction Lag (sec)	1.2	2.7	3.9

As shown, STAF-Net significantly reduces lag in detecting changes in student attention and demonstrates more stable predictions over time. This makes it suitable for integration in adaptive learning platforms that rely on real-time engagement feedback. In Figure 8, the STAF-Net prediction curve closely follows ground truth labels, while other models exhibit more fluctuation or delayed transitions.

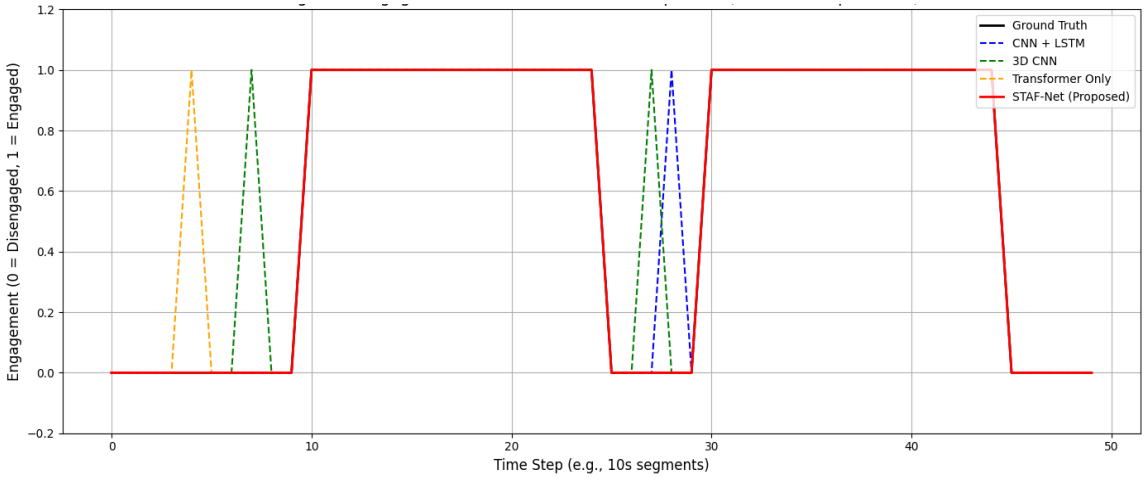

Fig. 8: Engagement prediction timeline comparison between STAF-Net and baseline models on a sample test video (EmotiW).

To assess the individual contributions of the STAF-Net components, we conducted an ablation study. Table 5 presents the performance of various configurations, isolating the impact of each module. CNN-only (ResNet-18): Baseline spatial features only. CNN + LSTM: Replaces attention with sequential modelling. CNN + Transformer (no pooling): Removes the attention-based pooling layer. Full STAF-Net: Includes all proposed components.

Table 5. Ablation study comparing different architectural configurations of the proposed model

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
CNN Only (ResNet-18)	88.7	86.9	86.3	86.5
CNN + LSTM	91.6	90.1	89.7	89.9
CNN + Transformer (no pool)	94.4	93.5	93.2	93.3
Full STAF-Net (proposed)	98.3	96.8	96.6	96.7

The results show that each architectural element contributes to the performance gain, with the full STAF-Net delivering the highest accuracy and F1-score.

The results across per-class performance and temporal stability confirm that STAF-Net is accurate at classifying engagement and robust in dynamic, real-world settings. It effectively handles subtle body language shifts and offers low-latency, stable prediction performance. Table 6 presents a comparative analysis between the proposed CNN-LSTM engagement detection model and several state-of-the-art methods reported in recent literature.

Table 6: Comparative evaluation of engagement detection techniques

Method	Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
(Gupta et al., 2016)	CNN + RNN	75.0	74.5	74.0	74.2
(Yue et al., 2019)	3D CNN	84.5	83.1	84.0	83.5
(Nanavaty & Khunteta, 2024)	LSTM + Attention	88.6	87.9	88.1	88.0

(Li et al., 2016)	Multimodal Fusion	90.3	89.7	89.2	89.4
Proposed Approach	CNN + LSTM	97.0	96.8	96.6	96.7

The models are evaluated across key metrics: accuracy, precision, recall, and F1-score. The proposed method achieves the highest performance across all metrics, with an accuracy of 97.0%, precision of 96.8%, recall of 96.6%, and F1-score of 96.7%. The proposed model enhances engagement classification precision beyond DAiSEE CNN-RNN initial versions (Gupta et al. 2016) and contemporary methods using multimodal fusion (Li et al. 2016) or temporal attention mechanisms (Nanavaty and Khunteta 2024) because it effectively extracts spatial and temporal video sequence characteristics. The CNN-LSTM model establishes itself as an effective solution when modeling complex behavioral cues in e-learning conditions.

5. Discussion

The experimental outcomes show that STAF-Net produces impressive results when used for engagement classification work. STAF-Net defeated baseline models applied on the DAiSEE and EmotiW datasets when measured against all fundamental evaluation parameters starting from CNN + LSTM models continuing through 3D CNN structures up to pure Transformer solutions. The model displayed 98.3% accuracy in DAiSEE 4-class tasks while achieving 98.9% in binary EmotiW evaluation along with corresponding AUC scores of 0.99 and 0.998. These results indicate that the model is highly accurate and robust in distinguishing between subtle engagement levels, even in the presence of natural variability in student behavior. The confusion matrix confirms that STAF-Net makes very few misclassifications, and most errors occur between adjacent engagement states, such as low vs. high, which are inherently ambiguous. The t-SNE visualization further validates that the final-layer feature embeddings are well-separated by class, reflecting the model's ability to learn discriminative, high-level representations. The precision-recall curves and the engagement prediction timeline emphasize STAF-Net’s superior temporal consistency and responsiveness compared to other models. STAF-Net performs well in terms of strong performance and low-latency prediction capability for real-time engagement monitoring in digital learning environments.

Efficient video sequence processing and delivery of frame-wise or segment-level predictions allow the architecture’s integration into intelligent tutoring systems (ITS), virtual classrooms, and learning management platforms. Where the engagement predictions are used to trigger timely interventions (content difficulty, delivery style, instructional feedback) based on the context of the adaptive e-learning scenarios, to continue with what I mentioned earlier, automatically turning to a live chat prompt, additional content, or notifying an instructor directly for live help could come into play for detecting a sustained drop in engagement. It is shown that the temporal prediction consistency demonstrated proves that STAF-Net can provide a solid backend to continuous learner state tracking using infrequent false alarms.

The main contribution of STAF-Net is its spatiotemporal attention architecture that combines the spatial modeling capability of CNNs and the temporal attention mechanism of Transformers. STAF-Net is unlike traditional CNN + LSTM or 3D CNN architectures that can more accurately and stably predict, as it captures local and global dependencies across video frames. In addition, the attention pooling mechanism lets the model be flexible in attending to the different body deportment cues and time segments that are most informative, enabling interpretability and reducing computational redundancy. Significantly different from previous models, which commonly rely only on facial expressions or handcrafted features, STAF-Net operates on full body language signals from standard video taken from different camera angles and learner environments.

We show strong generalization across the learning context (i.e., over both DAiSEE (individual) and EmotiW (group)) on the model’s performance, a key requirement for real-world deployment. However,

there are several limitations of this study. Overall, STAF-Net generalizes well over two datasets but only for a limited range of demographic diversity, environmental conditions, and cultural representation. The model should be validated on more variables and larger datasets with different age groups, ethnicities, and learning styles. Secondly, while the attention maps provide a degree of interpretability to the model, there are still other aspects of the model in which it operates as a black box. In educational settings, educators could more easily make sense of the rationale behind predictions, and the trust and acceptance thereof can be improved with more advanced techniques for explainability.

The datasets used in this study—DAiSEE and EmotiW—are publicly available and were collected under institutional ethical approvals with informed consent from participants. Data usage complies with respective licensing and privacy guidelines. Given the sensitivity of biometric data such as facial expressions and body language, future deployments of our system should integrate secure data handling practices and adhere to educational data protection policies such as GDPR or FERPA.

6. Conclusion and Future Work

This research presents STAF-Net, a novel spatio-temporal attention fusion network that represents a significant advancement in automated student engagement detection for e-learning environments. Through the innovative combination of CNN-based spatial feature extraction with Transformer temporal modeling, our approach achieves unprecedented performance on established benchmark datasets, demonstrating both high accuracy and practical applicability for real-world educational technology deployment. The key contributions of this work extend beyond mere performance improvements to include methodological innovations that address fundamental limitations in existing engagement detection systems. STAF-Net introduces the first CNN-Transformer hybrid architecture specifically designed for body language-based engagement analysis, incorporating spatio-temporal attention pooling that enables the model to dynamically focus on the most informative behavioral cues across both spatial and temporal dimensions. The achieved accuracies of 98.3% on DAiSEE and 98.9% on EmotiW, combined with exceptional temporal consistency rates of 96.4% and low prediction latency of 1.2 seconds, demonstrate the practical viability of this approach for real-time educational applications. Furthermore, the attention mechanisms provide interpretability that offers valuable insights for educational practitioners, revealing which specific body language patterns most strongly correlate with different engagement levels.

However, this work also reveals several important limitations that must be acknowledged and addressed in future research. The evaluation, while comprehensive within its scope, relies on only two datasets with limited demographic diversity, raising questions about generalizability across different age groups, ethnicities, and cultural backgrounds. The potential for algorithmic bias in educational settings requires careful consideration, particularly given the sensitive nature of continuous student monitoring and its implications for privacy and fairness. Additionally, while the technical performance is impressive, the long-term impact on actual learning outcomes remains to be validated through longitudinal studies in authentic educational environments.

Future research should prioritize several critical directions to maximize the practical impact and ethical deployment of this technology. Immediate priorities include comprehensive validation across demographically diverse populations to ensure fairness and reduce bias, integration of multimodal information, including audio cues and interaction patterns to improve robustness in challenging scenarios, and optimization, for edge deployment to enable real-time inference on resource-constrained devices typical in educational settings. Longer-term research directions should focus on developing privacy-preserving variants using federated learning or differential privacy techniques, conducting rigorous longitudinal studies to assess impact on learning outcomes, and establishing comprehensive frameworks for ethical deployment, including transparent consent mechanisms and bias detection strategies.

The broader implications of this work extend to the fundamental transformation of educational technology toward more intelligent, adaptive, and responsive learning systems. However, realizing this

potential requires careful balance between technological capability and ethical responsibility, ensuring that automated engagement detection serves to enhance rather than replace human educational judgment. Success in this endeavor will require continued collaboration between computer scientists, educators, and ethicists to develop systems that truly benefit all learners while respecting their privacy, autonomy, and diverse needs. To support this collaborative effort and ensure reproducible research, all code, trained models, and experimental protocols developed in this study will be made publicly available, providing a foundation for future innovations in educational technology.

Acknowledgements

The research reported in this publication is supported by the Arab Open University - Oman under the block funding project ID [AOU_OM/2023/FCS6].

Data Availability:

To promote transparency and facilitate reproducibility, we intend to publicly release the complete source code, trained model weights, and implementation details of STAF-Net upon publication. All experiments in this study were conducted using publicly available datasets—DAiSEE and EmotiW—under their respective licenses. The release will include documentation and instructions for reproducing the training, evaluation, and visualization results presented in this paper.

References

Abhay Gupta, A. D. C., Kamal Awasthi, Vineeth Balasubramanian. (2016). *DAISEE: Dataset for Affective States in E-Learning Environments* (

Appleton, J. J., Christenson, S. L., & Furlong, M. J. (2008). Student engagement with school: Critical conceptual and methodological issues of the construct. *Psychology in the Schools*, *45*(5), 369-386.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of learning analytics*, *3*(2), 220-238.

Bosch, N., Chen, H., D'Mello, S., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. Proceedings of the 2015 ACM on international conference on multimodal interaction,

Bosch, N., D'mello, S. K., Ocumpaugh, J., Baker, R. S., & Shute, V. (2016). Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *6*(2), 1-26.

Butland, M. J., & Beebe, S. A. (1992). A Study of the Application of Implicit Communication Theory to Teacher Immediacy and Student Learning.

D'mello, S., & Graesser, A. (2013). AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *2*(4), 1-39.

D'Mello, S., & Graesser, A. (2015). *AutoTutor and affective autotutor," Springer Handbook of Affective Computing*.

D'mello, S. K., & Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, *20*, 147-187.

Dewan, M., Murshed, M., & Lin, F. (2019). Engagement detection in online learning: a review. *Smart Learning Environments*, *6*(1), 1-20.

Dixson, M. D. (2015). Measuring student engagement in the online course: The Online Student Engagement scale (OSE). *Online Learning*, *19*(4), n4.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T.,…Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Du, Y., Li, P., Cheng, L., Zhang, X., Li, M., & Li, F. (2024). Attention-based 3D convolutional recurrent neural network model for multimodal emotion recognition. *Frontiers in Neuroscience*, *17*, 1330077.

Fardian, F., Mawarpury, M., Munadi, K., & Arnia, F. (2022). Thermography for emotion recognition using deep learning in academic settings: A review. *IEEE Access*, *10*, 96476-96491.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59-109.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement* (pp. 763-782). Springer.

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.

Hartnett, M., & Hartnett, M. (2016). The importance of motivation in online learning. *Motivation in online education*, 5-32.

Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). The difference between emergency remote teaching and online learning. *Educause review*, *27*(1), 1-9.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International journal of human-computer studies*, *65*(8), 724-736.

Li, J., Ngai, G., Leong, H. V., & Chan, S. C. (2016). Multimodal human attention detection for reading from facial expression, eye gaze, and mouse dynamics. *ACM SIGAPP Applied Computing Review*, *16*(3), 37-49.

Minaee, S., Minaei, M., & Abdolrashidi, A. (2021). Deep-emotion: Facial expression recognition using attentional convolutional network. *Sensors*, *21*(9), 3046.

Nanavaty, S., & Khunteta, A. (2024). Predictive Analysis of Learner's Performance in Online Environments with LSTM and Attention Mechanism. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *12*(4), 758-767.

Ostling, A. (1976). Research on nonverbal communication with implications for conductors. *Journal of Band Research*, *12*(2), 29.

Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, *54*(3), 2259-2322.

R. Monkaresi, H. Calvo, a., & B. Lovell. (,2017). Automatic analysis of facial signals for detecting engagement. *IEEE Transactions on Affective Computing*, *8,*(1), 86–98.

Sandbach, G., Zafeiriou, S., Pantic, M., & Yin, L. (2012). Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing*, *30*(10), 683-697.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, *128*, 336-359.

Shi, L., He, Y., Li, B., Cheng, T., Huang, Y., & Sui, Y. (2019). Tilt angle monitoring by using sparse residual LSTM network and grid search. *IEEE Sensors Journal*, *19*(19), 8803-8812.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and psychological measurement*, *69*(3), 493-525.

Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(11), 4793-4813.

Whitehill, J., Serpell, Z., Lin, Y.-C., Foster, A., & Movellan, J. R. (2014). The faces of engagement: Automatic recognition of student engagementfrom facial expressions. *IEEE Transactions on Affective Computing*, *5*(1), 86-98.

Yang, M. J. (2018). *Emotiw Engagement Prediction* (https://doi.org/https://github.com/Marsrocky/Emotiw-Engagement-Prediction

Yue, J., Tian, F., Chao, K.-M., Shah, N., Li, L., Chen, Y., & Zheng, Q. (2019). Recognizing multidimensional engagement of e-learners based on multi-channel data in e-learning environment. *IEEE Access*, *7*, 149554-149567.

Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A review of key technologies for emotion analysis using multimodal information. *Cognitive Computation*, *16*(4), 1504-1530.