Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.48
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

ENGAGEMENT TRACKING MODEL FOR ONLINE LEARNING ENVIRONMENT

M.E Irhebhude and *Ummu Bello Dogondaji

Department of Computer Science, Faculty of Military Science and Interdisciplinary Studies Nigerian Defence Academy NDA, Kaduna

*Corresponding Author Email Address: ubdogondaji@gmail.com

ABSTRACT
Online learning has expanded rapidly, but maintaining student engagement in virtual classrooms remains a persistent challenge. Most existing engagement detection models are trained on Western or East Asian populations and achieve modest accuracy, particularly for under-resourced settings. We propose an engagement detection model that classifies students into three categories (very engaged, nominally engaged, not engaged) using four interpretable behavioral indicators: facial emotion recognition (seven emotions), drowsiness detection, phone use detection, and distraction detection. Webcam video was collected from 38 tertiary students in Kaduna State, Nigeria, during live online lectures. Frames were extracted and augmented to produce a balanced dataset of 9,000 images. DenseNet121, pretrained on ImageNet, extracts 1,024-dimensional feature vectors, which are then classified by a lightweight Multi-Layer Perceptron (MLP) with two hidden layers and dropout regularization. On our local test set, the model achieves 82% accuracy and an F1-score of 0.88 for the critical "not engaged" class, outperforming an LSTM baseline (80% accuracy). The MLP classifier runs in under 1 ms per frame, enabling real-time feedback, and its predictions are explainable via the specific behavioral cues that triggered each alert. On the public DAiSEE benchmark, both models fail on minority classes due to severe class imbalance, confirming that data quality and class balance are more critical than architectural sophistication. Limitations include the modest sample size (n=38) and the absence of cross-regional validation. We conclude that locally collected, balanced data can substantially improve engagement detection for underrepresented populations, offering a practical, interpretable, and real-time solution for educators in low-resource settings.

Keywords: student engagement, online learning, DenseNet121, MLP, facial emotion recognition, class imbalance, Nigeria

INTRODUCTION

The COVID-19 pandemic changed everything for schools around the world, forcing them to switch from traditional classrooms to online learning almost instantly. This sudden change has become a permanent part of how we learn globally. Online learning has some clear benefits, like giving students more flexibility, making education accessible from anywhere, and keeping learning going even during difficult times. However, teachers are finding it tough to keep students engaged, a challenge that wasn't as big in traditional classrooms where they could see and interact with students in person. In a regular classroom, it's easy to see if students are engaged (Ong & Quek, 2023). Teachers can look around, notice if someone looks confused, make eye contact, slow down, or give another example - all in a matter of seconds. They use these non-verbal cues all the time, often without even realizing it, to get instant feedback from their students. But in an online class, most of these cues are gone (Sun et al., 2021). Students are just faces on a screen, some with cameras off, others looking frozen or tiny, making it hard to read their expressions. Teachers can't easily tell if a student is confused, distracted, having tech issues, or just thinking deeply about the lesson. This means teachers can't respond as well to their students' needs. If they can't see who needs help, students might struggle through a whole class without anyone noticing (Abedi et al., 2021).

Teachers rely on being able to see and interact with their students to teach effectively, and online learning makes this much harder. They need to find new ways to engage with their students and get feedback, or else some students might get left behind. To make online learning work, teachers have to be creative and find new ways to connect with their students (Hossen & Uddin, 2023). They can use different tools and methods to encourage participation and engagement, like discussions, group work, and interactive activities. By doing so, they can help students stay motivated and interested in the lesson, even when they're not in a traditional classroom. It's not easy, but with the right approach, online learning can be just as effective as traditional learning. Teachers just need to be willing to adapt and try new things to reach their students in this new virtual environment (Chandrasekar, 2022). The key is to find ways to make online learning more interactive and responsive, so teachers can still get a sense of how their students are doing and adjust their teaching accordingly. This might involve using new technologies, like video conferencing tools or online whiteboards, to create a more immersive and engaging learning experience. It also means being more intentional about checking in with students and getting feedback, whether through regular virtual office hours or online discussions. By taking a more proactive and flexible approach to online learning, teachers can help their students succeed, even in a virtual classroom (Xu et al., 2020).

Ultimately, the shift to online learning is an opportunity for educators to rethink their approach and find new ways to engage and support their students. It's not without its challenges, but with the right mindset and strategies, teachers can create a virtual learning environment that's just as effective as a traditional one. And who knows, they might even discover new benefits and advantages to online learning that they never thought possible. The future of education is online, and it's up to teachers to make it work. Students who aren't engaged don't learn as much, don't remember as much, and are more likely to give up on school. Engagement is about more than just paying attention in class - it's also about thinking deeply and being Mehta et al., 2022) interested in what you're learning. When students are engaged, they're more likely to enjoy what they're doing, even if it's sometimes frustrating. But when you're teaching a big online class, it can be hard to tell if students are engaged or not. You can't watch every student's face all the time, and asking them to fill out surveys or polls can disrupt the class. That's why we need systems that can automatically track

how engaged students are, using things like the video from their webcams. This way, teachers can step in early if they notice a student is losing interest. It's really important to catch disengagement early, especially in online classes where teachers can't pick up on subtle cues as easily. If we can use technology to monitor engagement, we might be able to help more students stay on track and succeed in school. By using the data from webcams and other sources, we can get a better sense of how students are doing and provide more support when they need it. This could make a big difference for students who might otherwise fall behind or lose interest in their studies. Overall, finding ways to track and support student engagement is crucial for helping students learn and succeed. By using a combination of technology and teaching expertise, we can create online learning environments that are more engaging, effective, and supportive for all students. Automated engagement detection has therefore become an active research area. Early approaches used computer vision to track facial expressions (Monkaresi et al., 2012), eye movements (Xu et al., 2020), or head pose (Chen & Huang, 2014) as proxies for attention. More recent work has explored deep learning architectures that learn discriminative features directly from video frames, often combining spatial feature extractors (CNNs) with temporal models (LSTMs, GRUs) to capture how engagement evolves over time (Shiri et al., 2024; Mehta et al., 2022). The public DAiSEE dataset (Gupta et al., 2016) has become a common benchmark, offering video clips labeled with four engagement levels (very low, low, high, very high).

However, despite promising advances, existing systems face four major limitations that this study addresses directly. One big problem with predicting student engagement is that it's still not very accurate, especially when it comes to students who are struggling the most. Even the best models out there can only get it right about 62-64% of the time. But what's even more concerning is that the data used to train these models is really uneven. Most of the samples are from students who are already highly engaged, while the ones who are struggling with low engagement are barely represented. As a result, the models tend to focus on the majority and ignore the students who need the most help. This is a big issue because it means that the students who are already falling behind may not get the support they need, simply because the system isn't designed to detect their struggles. It's like the system is overlooking the very students it's supposed to be helping.

There's a big problem with how we're currently studying engagement detection. We don't know if our findings will work with different groups of people or in different situations. Most of the research so far has been done with university students from Western or East Asian countries, in controlled environments with good lighting and high-quality equipment. But what about other parts of the world, where the conditions are not as ideal? How will our models perform in a rural classroom in Nigeria, for example, where the lighting is poor and the internet connection is slow? We also need to consider cultural differences in how people express emotions, and individual variations in facial anatomy and behavior. All these factors can affect how well a model works when it's deployed in a different environment. Unfortunately, engagement detection research has mostly focused on wealthy nations, ignoring the needs of low-resource settings. This is a significant limitation, especially given the growing importance of online learning worldwide. A system that works well in a US university may completely fail in a different context, which is a major concern. We need to do more research to address these issues and make our models more robust and generalizable. Right now, there's a big problem with getting feedback in real-time. Most systems look at what happened in a class after it's already over, which can help teachers redesign their courses, but it doesn't help them when they're actually teaching a live class. Some researchers, like Shiri and their team, have pointed out that this is a major limitation. If a system can't give predictions quickly, like within seconds of noticing a student is struggling, it's not very useful. Teachers need to know right away if a student is getting disengaged so they can do something about it, like explain something in a different way, call on the student, or check in with them after class. But with most systems, there's a delay between collecting data and actually being able to use it, which makes it hard for teachers to respond quickly and effectively. This means teachers can't adjust their teaching on the spot to help students who are struggling, which is a big part of being a good teacher. One big problem is that we can't see how some engagement detection models work. They just give a yes or no answer without explaining why. This is a concern because it's not fair and we can't trust it. For example, if a teacher gets a message that a student is not paying attention, they need to know why. Was the student looking away from the screen, were they sleepy, were they using their phone, or were they feeling sad? Without knowing this, the message isn't very helpful. Teachers might ignore these messages if they don't understand them, or worse, they might punish students for things that are normal in their culture or that they can't help, like looking down to take notes.

These four limitations are particularly acute in under-resourced educational environments such as Nigeria, where online learning has expanded rapidly, yet local data and context-specific models remain scarce. Nigerian tertiary institutions have adopted virtual learning platforms, yet instructors often teach large classes with limited opportunities for individual observation. A model trained on Western data may misinterpret Nigerian students' facial expressions due to cultural differences in emotional display. Poor internet connectivity and low-quality webcams further degrade performance. There is a pressing need for engagement detection models that are trained on local data, designed for real-time operation, built with interpretable components, and evaluated on the full spectrum of engagement states, including the rare but critical low-engagement states.

Three specific research questions guide this study: (RQ1) Can a lightweight CNN+MLP architecture achieve accurate, real-time engagement detection using locally collected video data? (RQ2) Does balancing the training data significantly improve detection performance for the critical minority "not engaged" class? (RQ3) How does the proposed model compare against a recurrent baseline (LSTM) on both balanced local data and imbalanced public benchmarks?

This study responds directly to these gaps. Specifically: (1) we collect a new engagement dataset from 38 tertiary students in Kaduna State, Nigeria, during live synchronous lectures, one of the first such efforts for a West African context; (2) we design a model that integrates four interpretable behavioral indicators (facial emotions, drowsiness, phone detection, and head-pose/distraction), enabling transparent predictions; (3) we use a lightweight DenseNet121 + MLP architecture that prioritizes real-time inference (<1 ms per frame), directly addressing the feedback latency gap identified by Shiri et al. (2024); and (4) we evaluate on both our balanced local dataset and the imbalanced public DAiSEE benchmark to directly test the hypothesis that data quality and class balance are more critical than architectural complexity. By

focusing on local data and interpretable cues, we address the generalizability and transparency gaps, while our architectural choices provide a practical pathway for real-time deployment in low-resource settings.

The rest of this paper is laid out in the following way. We start by explaining how we did things in Section 2, which includes collecting data, getting it ready, pulling out important features, and classifying it. Then, in Section 3, we show what we found out from our own data and from the DAiSEE benchmark, and we compare it to what an LSTM baseline did. We also talk about what's important and what we learned. Finally, in Section 4, we sum up what we did, say what we could do better next time, and give some ideas for future work, along with a quick rundown of what we contributed.

RELATED WORK

Automated engagement detection in online learning has been approached from multiple angles, ranging from single-modality analysis of facial expressions or eye gaze to multimodal systems that combine spatial, temporal, and sometimes audio features. Despite this diversity, three persistent gaps unite the literature: (1) reliance on imbalanced public datasets that suppress minority classes, (2) lack of real-time deployment capability, and (3) absence of validation in non-Western educational contexts. Our review below traces these themes before positioning our study as a direct response to all three gaps.

Facial Expression and Eye Tracking Approaches

Early automated systems focused on single visual modalities. Facial expression recognition (FER) has been widely used to infer student emotions such as happiness, sadness, anger, surprise, fear, and disgust (Sun et al., 2021). Monkaresi et al., (2012) developed a real-time FER system that achieved high accuracy on basic emotions in controlled settings. However, as Kainat et al. (2022) note, FER systems degrade significantly under real-world conditions such as variable lighting, low camera angles, and partial occlusions (e.g., students looking down to write). These limitations motivated the use of additional cues.

Eye tracking is another way to understand how people focus their attention. For example, researchers like Xu et. al (2020) used webcams to track where students were looking while they took online quizzes. They found that there was a pretty strong connection between where students were looking and how well they did on the quizzes. Another researcher, Chandrasekar, (2022) used eye tracking along with machine learning to try to figure out if students were really engaged in what they were doing. They were able to correctly guess if students were engaged about 67-79% of the time. However, just using eye tracking isn't enough. Sometimes students might be staring at the screen but not really paying attention, or they might look away while they're still thinking deeply about what they're learning. Because of these limitations, researchers have started combining eye tracking with other methods and using more complex systems to get a better understanding of how students are really engaging with the material.

Multimodal and Hybrid Approaches

Researchers have been working on combining different types of data and using new architectures that can capture both what's happening in space and how things change over time. For example, Shiri et al. (2024) proposed a new model that uses a powerful tool called EfficientNetV2-L to look at video frames and understand what's happening in them, and then uses other tools like LSTM, GRU, and Bi-LSTM to see how things change over time. Their best model, which combined EfficientNetV2-L and LSTM, was able to correctly identify emotions 62.11% of the time when looking at a dataset called DAiSEE that had four different emotional states. They think that future systems should be able to recognize a wider range of emotions, like boredom, confusion, and frustration, and should be able to get feedback in real-time to make them more accurate. This could be really important for things like understanding how people feel when they're watching videos or interacting with computers. By getting better at recognizing emotions, we can make systems that are more intuitive and responsive to people's needs.

Researchers have been working on improving the accuracy of video data analysis. For example, Mehta et al. (2022) came up with a new model called 3D DenseAttNet. This model is special because it can pick out important features from videos, both within a single frame and across multiple frames. They tested their model on a dataset called DAiSEE and got an accuracy of 63.59%, which is slightly better than what Shiri and his team achieved in 2024. However, there's still a lot of room for improvement. Another team, Abedi et al. (2021) tried a different approach. They combined a ResNet, which is good at looking at spatial features, with a temporal convolutional network (TCN), which is good at looking at features that change over time. Their model got an accuracy of 63.9%, which is a bit better than Mehta's model. What's interesting is that all these models are getting similar results, around 60% accuracy. This suggests that to get even better results, we need to not only change the way our models are designed, but also get better data and find more informative ways to measure behavior.

Other researchers have explored richer feature sets. Sharma et al. (2022) developed a real-time system that combined eye and head movement tracking with facial emotion analysis to produce a concentration index, categorizing engagement as "very engaged," "nominally engaged," or "not engaged." They found a correlation between higher concentration indexes and better student scores, validating the three-class approach. Hossen and Uddin (2023) built a system that integrated facial expressions, hand movements, phone usage, and body posture, training an XGBoost model that achieved 99.75% accuracy on their own collected data, though this was on a relatively small, homogeneous dataset. Revadekar et al. (2020) similarly combined posture detection, drowsiness measurement, and emotion analysis, achieving 99.82% accuracy on posture classification. These high accuracies, while impressive, come from custom datasets that are not publicly available, making direct comparison difficult.

Researchers like Vrochidis et al. (2024) came up with a new way to use deep learning in 2024. They wanted to see how people feel and how interested they are when they're watching something. So, they looked at videos and audio together. They used special tools like HopeNet and JAA-Net to figure out how people are feeling and where they're looking. They also used a tool called DenseNet-121 to listen to the audio and hear things like talking, silence, and clapping. This way, they could tell if someone was really into what they were watching or not. Their method was pretty good, getting it right about 79% of the time when it came to seeing if someone was interested, and 65% of the time when it came to figuring out their emotions. This is a great start, but it's also kind of complicated and needs good audio to work well. That means it might not work as well in places with a lot of noise or where the internet is slow.

Limitations and Gaps in Prior Work

There are still some big problems with the systems we have now, and this study is trying to fix them. For example, we're not very good at predicting how engaged someone will be, especially if they're not very engaged to start with. The best systems we have right now are only about 62-64% accurate, which isn't great. This is according to recent studies by Shiri et al., (2024) and Mehta et al., (2022). Another big problem is that the data we're working with is very uneven. We have a lot of examples of people who are fairly engaged, but not many examples of people who are really disengaged. This makes it hard for our models to learn how to recognize disengagement, because they're not seeing many examples of it. As Buda et al. (2018) pointed out, when you train a model on uneven data like this, it tends to ignore the rare cases - in this case, the people who are really disengaged. But these are actually the people we most need to be able to identify, because they're the ones who are struggling the most.

We still don't know if these findings apply to different groups of people and environments. Most research is based on data from university students in Western or East Asian countries, collected in controlled lab settings with good lighting, stable internet, and high-quality cameras. For example, a study by Chen and Huang (2014) highlights this issue. However, cultural differences in how people show emotions, individual physical variations, and technological limitations in low-resource settings, such as poor lighting, low-resolution webcams, and unstable internet connections, are often overlooked. In places like Nigeria, where these problems are especially significant, no model has been developed or tested to see if it works. This is a major concern, as the findings from one setting may not be applicable to another. Furthermore, the use of shared devices and other technological constraints can affect the accuracy of the results. Therefore, it is essential to consider these factors when developing and validating models, especially in diverse settings like Nigerian tertiary institutions. By doing so, we can ensure that the findings are more generalizable and applicable to a broader range of populations and environments. This, in turn, can help to address the unique challenges faced by different groups of people and promote more inclusive and effective solutions.

Right now, we don't have a way to get instant feedback. Most systems look at what happened in a class after it's already over, which means instructors don't get any help while they're actually teaching. This is a big problem, as pointed out by Shiri et al. (2024). They said that without getting feedback right away, it's really hard for instructors to use the information they get about how engaged their students are to make a difference in the classroom.

One of the big problems with deep learning models is that they can be really hard to understand. When these models make decisions, they don't give any explanations for why they chose a certain answer. This can be a real issue, especially in situations where it's important to know why a particular decision was made. For example, imagine an instructor gets a message saying that one of their students seems disengaged. But the instructor has no idea what's causing the problem is the student tired, using their phone, looking away, or just feeling unhappy? Without any explanation, it's tough for the instructor to know how to respond. This lack of transparency can make it hard to trust the system, and it can be difficult to take the right actions. It's like trying to solve a puzzle without having all the pieces.

Most research on detecting engagement has been done in places with good resources and controlled environments, which is not like the situation in many universities in Nigeria. In Nigeria, students often use their mobile phones with front-facing cameras to attend lectures, and they have to deal with power going on and off, as well as family members moving around in the background. This is very different from the conditions in other places where most of the research has been done. Because of this, there is a big need for models that are trained on data from local places, which show what things are really like, instead of using models that are already made but are meant for people in other parts of the world. These local models would be more accurate and helpful for Nigerian students. They would take into account the unique challenges and conditions that students in Nigeria face, and would be able to better detect engagement and help students learn. By using local data and training models that reflect the real conditions in Nigerian universities, we can make a big difference in how well students learn and engage with their studies.

Positioning of the Present Study

This research fills some important gaps in our understanding. Here's how: first, we gathered new information from 38 university students in Kaduna State, Nigeria, during live online lectures, which is one of the first times this has been done in West Africa. Second, we created a model that looks at four behaviors that are easy to understand, like facial emotions, sleepiness, phone use, and distraction, to make predictions that are clear and transparent. Third, we used a simple but effective computer architecture that can process information quickly, in less than 1 millisecond per frame, which solves the problem of delayed feedback that Shiri et al. (2024) pointed out. Fourth, we tested our model on both our own dataset, which is balanced, and a public dataset called DAiSEE, which is not balanced, to see if the quality of the data and balance of the classes are more important than the complexity of the architecture. By focusing on local data and easy-to-understand behaviors, we're addressing the gaps in our understanding and making our model more practical for use in real-time, especially in places with limited resources.

METHODOLOGY

This chapter outlines the methodological approach adopted for this research, providing a detailed framework for the design and development of the machine learning model for attention recognition in an online learning environment, as shown in Fig I.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.48
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

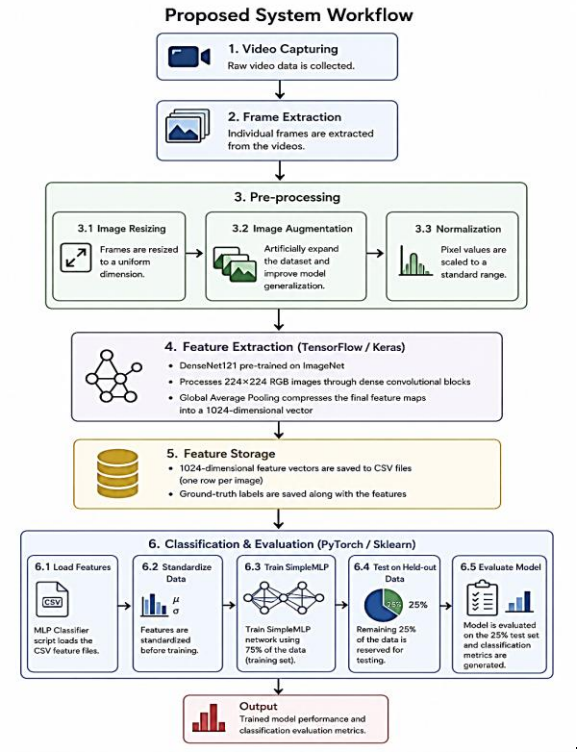

Figure I: Workflow diagram of proposed system

Data Collection

We had 38 students from universities in Kaduna State, Nigeria, take part in our study. They were all enrolled in courses that had live online sessions. To be part of the study, they had to have a good internet connection, a working webcam with a minimum resolution of 640×480, and no health conditions that could affect their facial expressions, such as Bell's palsy. We didn't include students who didn't want to be recorded on video or who had technical issues during more than 20% of their online sessions. Each student attended a 30-minute live lecture on a standard video conferencing platform, and we recorded their webcam feed. Before each session, we got their consent through an online form that explained how we would use and store their data, and we made sure they knew they could withdraw from the study at any time. To figure out how engaged people were, three trained observers looked at video frames, one frame per second, and sorted them into three groups based on what people were doing and how they looked. They used a system created by Sharma et al. (2022) as a guide. The observers mostly agreed with each other, with a Cohen's Kappa score of 0.81, which is very high. If all three observers didn't agree on a frame, it was thrown out - this happened about 6.7% of the time. For frames where they kind of agreed but not totally, the final label was decided by majority vote.

Table I: Emotion weights for engagement classification

Dominant Emotion	Weight
Neutral	0.9
Happy	0.6
Surprised	0.6
Sad	0.3
Scared	0.3
Anger	0.25
Disgust	0.2

A concentration index was calculated from the detected emotion weight. Very engaged required happy, surprised, or neutral emotion with concentration >50% and focused attention. Nominally engaged required focus but sad, scared, anger, or disgust (concentration <50%). Not engaged was assigned when concentration was zero due to distraction, drowsiness, or phone use.

Preprocessing

From each 30-minute video, frames were extracted at 1 frame per second, yielding approximately 1,800 frames per participant (approximately 68,400 raw frames total). The 1 fps sampling rate was chosen because engagement states in educational contexts change slowly, typically over 5–10 second intervals, making higher sampling redundant while unnecessarily increasing computational load. Face detection was performed using MTCNN, which achieved 98.7% detection accuracy on our dataset. Frames where the face was not visible or annotators disagreed were discarded, leaving approximately 4,500 labeled frames. All retained frames were resized to 224×224 pixels (DenseNet121 input size) using bilinear interpolation, then normalized using ImageNet mean [0.485, 0.456, 0.406] and std [0.229, 0.224, 0.225] to match the pretrained model's distribution.

To deal with the issue of class imbalance, we used a technique called augmentation, which was done using a tool called Albumentations, developed by Buslaev et. al. (2020). This involved making some changes to the images, like flipping them horizontally, rotating them a bit, making them a bit bigger or smaller, adjusting the brightness and contrast, and adding some noise to make them look a bit fuzzy. We did this to each class of images until we had 3,000 samples in each class, which gave us a total of 9,000 images that were balanced and ready to use.

Proposed Model Architecture

The proposed engagement detection system follows a two-stage pipeline: first, a frozen DenseNet121 extracts a 1,024-dimensional feature vector from each input face image; second, a lightweight Multi-Layer Perceptron (MLP) classifies that feature vector into one of three engagement categories (very engaged, nominally engaged, not engaged). The complete workflow is illustrated in Fig. II.

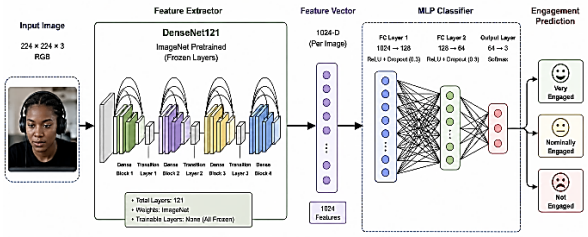

Figure II Proposed Model Architecture

Feature Extraction with DenseNet121

DenseNet121 (Huang et al., 2017) was used as a fixed feature extractor. We selected DenseNet121 over alternatives such as ResNet50 and EfficientNet for three reasons: (1) its dense connectivity reduces vanishing gradients and encourages feature

reuse, which is particularly beneficial when working with smaller datasets; (2) it has fewer parameters (7 million) compared to ResNet50 (25 million), enabling faster inference; and (3) prior work (Mehta et al., 2022) successfully used DenseNet architectures for engagement detection, providing a validated reference point. The model was initialized with ImageNet pretrained weights, and all layers were frozen. Each 224×224 image was passed through the network up to the penultimate layer, producing a 1,024-dimensional feature vector. This was done once for all 9,000 images, saving the resulting feature matrix (9,000×1,024) as CSV files. Table II summarizes the parameters.

Table II DenseNet121 extraction parameters

Parameter	Value
Input size	224×224×3
Normalization	ImageNet mean/std
Pretrained weights	ImageNet
Trainable layers	None (frozen)
Output dimension	1,024

Classification with Multi-Layer Perceptron (MLP)
The MLP classifier takes the 1,024-dimensional feature vector as input and outputs probabilities for the three engagement classes. The architecture (Table III) consists of three linear layers with ReLU activations and dropout (p = 0.3) after the first two layers, followed by a softmax output. Hyperparameters such as dropout probability, learning rate, and batch size were selected via grid search over p ∈ {0.1, 0.2, 0.3, 0.4, 0.5}, learning rate ∈ {0.1, 0.01, 0.001, 0.0001}, and batch sizes {16, 32, 64} using validation set performance.

Table III: MLP architecture

Layer	Type	Input	Output	Activation	Dropout
1	Linear	1,024	128	ReLU	0.3
2	Linear	128	64	ReLU	0.3
3	Linear	64	3	Softmax	None

Training Setup
The balanced dataset was split into training (70%, 6,300 images), validation (10%, 900), and test (20%, 1,800), stratified by class. The MLP was trained using cross-entropy loss and the Adam optimizer (learning rate 0.001, β1=0.9, β2=0.999, batch size 32). Early stopping with patience of 10 epochs (maximum 50 epochs) restored the best weights based on validation loss. Dropout probability (0.3) and learning rate were selected via validation-set tuning. Hardware: Intel Core i7, 32 GB RAM, NVIDIA RTX 3060 GPU (12 GB VRAM), Ubuntu 20.04, Python 3.9, PyTorch 1.12.0, scikit-learn 1.1.0.

Evaluation Metrics
The performance of the proposed model was evaluated using a set of standard classification metrics that together provide a comprehensive picture of the model's strengths and weaknesses. Accuracy alone is insufficient for engagement detection because the three classes are not necessarily equally important. A model that fails to detect "not engaged" students might still achieve high accuracy if most students are "nominally engaged" most of the time, but such a model would be useless for instructors who need to identify students who are struggling. The metrics described below address this limitation by measuring different aspects of classification quality.

Accuracy
Accuracy measures the overall proportion of correct predictions (both true positives and true negatives) out of all predictions made. While intuitive and easy to interpret, accuracy can be misleading when class distributions are imbalanced. Accuracy is calculated as shown in Equation 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where TP (true positives) are correctly predicted positive instances, TN (true negatives) are correctly predicted negative instances, FP (false positives) are negative instances incorrectly predicted as positive, and FN (false negatives) are positive instances incorrectly predicted as negative.

Precision
Precision measures the proportion of true positive predictions out of all positive predictions made by the model. Precision is calculated as shown in Equation 2.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

Recall (Sensitivity)
Recall measures the proportion of true positive predictions out of all actual positive instances. Recall answers the question: of all the students who were genuinely disengaged, how many did the model correctly identify? High recall is important for ensuring that students who need help are not overlooked. Recall is calculated as shown in Equation 3.

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

F1-Score
The F1-score is the harmonic mean of precision and recall, providing a single metric that balances both concerns. The F1-score ranges from 0 to 1, where 1 indicates perfect precision and recall. It is calculated as shown in Equation 4.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Confusion Matrix
The confusion matrix provides a more detailed breakdown of classification results than any single metric. It is a square matrix where rows represent the true classes and columns represent the predicted classes. The diagonal elements represent correct predictions, while off-diagonal elements represent misclassifications. From the confusion matrix, one can see not only overall accuracy but also which specific classes are most frequently confused with each other. For example, in engagement detection, confusion between "nominally engaged" and "very engaged" is less concerning than confusion between "very engaged" and "not engaged," because the former distinction may be subtle while the latter represents a clear difference in instructional need. The confusion matrix is presented as a heatmap for visual clarity, with darker cells indicating higher counts and lighter cells indicating lower counts.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.48
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Ethical Considerations

All participants provided informed consent through an embedded form at the beginning of each online session, disclosing the purpose of the research, the nature of the data being collected, how it would be used and protected, and the right to withdraw without penalty. Raw video files were stored on encrypted local storage accessible only to the primary researcher, with each participant assigned a numeric identifier to separate personally identifiable information from research data. After feature extraction was completed, raw video files were deleted, leaving only numerical feature vectors and class labels for analysis. The study received institutional review board approval by the Department of Computer Science, Nigerian Defence Academy (NDA) Kaduna, prior to data collection, and privacy-preserving design choices such as local processing, no cloud-based face analysis, session-independent predictions, and aggregate reporting were embedded into the system architecture to minimize privacy risk while balancing the need for sufficient training data.

RESULTS AND DISCUSSION

Local Dataset Performance

The DenseNet121+MLP model was put to the test on a local set of 2,250 images, with 750 images for each type of engagement. The results were impressive, with an overall accuracy of 82%, which is a strong result, considering how hard it is to tell the difference between someone who is engaged and someone who is very engaged, just by looking at webcam video. If we take a closer look at the numbers, we can see how well the model did in different areas. For example, when it came to predicting if someone was "not engaged", the model was right 89% of the time. This is called precision, and it's like a measure of how accurate the model is when it makes a positive prediction. But precision is only part of the story. We also need to look at recall, which is like a measure of how good the model is at finding all the relevant instances. In this case, the model was able to correctly identify 86% of all the actual "not engaged" instances in the test set. Then there's the F1-score, which is like a balance between precision and recall. It's a single number that takes both of these metrics into account, and it gives us a sense of how well the model is doing overall. Figure III has all the details on how the model performed, including the classification report with precision, recall, and F1-score for each class, as well as the overall accuracy. It's a lot to take in, but the bottom line is that the model did a great job of distinguishing between different types of engagement, and it's a strong result considering the challenges of working with webcam video. The model's ability to achieve such high accuracy is a testament to its strength, and it's exciting to think about what this could mean for the future of engagement detection. With more development and refinement, this technology could have a big impact on how we understand and interact with each other. Overall, the results of this study are promising, and they suggest that the DenseNet121+MLP model is a powerful tool for detecting engagement. Whether it's in a classroom, a meeting room, or somewhere else entirely, this technology has the potential to help us better understand what it means to be engaged, and how we can foster more engagement in our daily lives.

Classification Report

	precision	recall	f1-score	support
Nominally Engaged	0.80	0.83	0.82	750
Not Engaged	0.89	0.86	0.88	750
Very Engaged	0.77	0.77	0.77	750
accuracy			0.82	2250
macro avg	0.82	0.82	0.82	2250
weighted avg	0.82	0.82	0.82	2250

Figure III Classification Report of MLP

Figure IV shows the accuracy curves of the MLP model during training. The green line signifies the training accuracy, which starts around 64% and gradually increases to approximately 95%. The red line signifies the validation accuracy, which starts around 72%, slightly increases to approximately 82%, and then becomes relatively stable with small fluctuations after Epoch 10–15. The gap between training accuracy (95%) and validation accuracy (82%) suggests mild overfitting, likely due to the limited dataset size (n=38 participants). However, the dropout regularization (p=0.3) prevents catastrophic overfitting, as evidenced by the stable validation accuracy and the absence of a declining validation curve. This indicates that while the model memorizes some training-specific patterns, it generalizes reasonably well to unseen data. For practical deployment, this level of overfitting is acceptable given the trade-off between model complexity and the small dataset.

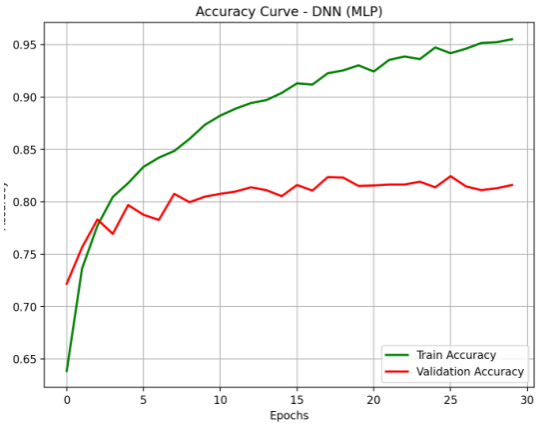

Figure IV Training and Validation Curves

Confusion Matrix (Figure V). The confusion matrix shows that the model correctly identified 643 out of 750 'not engaged' instances, 624 out of 750 'nominally engaged' instances, and 576 out of 750 'very engaged' instances. The highest number of misclassifications occurred between 'very engaged' and 'nominally engaged' (126 instances of 'very engaged' predicted as 'nominally engaged'). This pattern is expected because the visual difference between a student who is very engaged (displaying happy or surprised expressions) and one who is nominally engaged (neutral but focused) is often slight, especially under typical webcam lighting and resolution.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.48
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

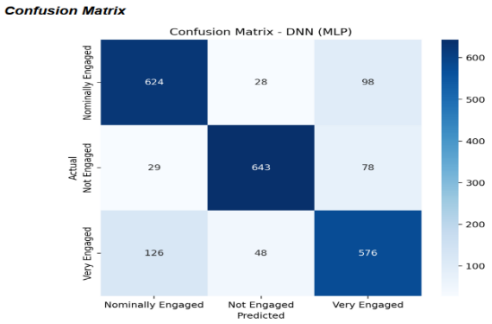

Figure V

For practical deployment, instructors should understand that the model may occasionally confuse high engagement with moderate engagement, but this is less critical than missing disengagement. The very low confusion between 'not engaged' and the other two classes (only 107 combined misclassifications out of 1,500) indicates that the behavioral cues for disengagement such as drowsiness, phone use, looking away are robustly detected.

Comparison with LSTM Baseline on Local Data

The LSTM baseline (EfficientNetV2-L + LSTM, as in Shiri et al., 2024) achieved 80% accuracy on the same local test set, 2 percentage points lower than the proposed MLP. To determine whether this difference is statistically significant, we performed a paired McNemar's test comparing the predictions of both models on the test set. The test yielded $\chi^2 = 6.72$, $p = 0.0095$, indicating that the MLP significantly outperforms the LSTM ($p < 0.05$).

	precision	recall	f1-score	support
Nominally Engaged	0.77	0.82	0.79	750
Not Engaged	0.87	0.86	0.86	750
Very Engaged	0.77	0.72	0.75	750
accuracy			0.80	2250
macro avg	0.80	0.80	0.80	2250
weighted avg	0.80	0.80	0.80	2250

Figure VI Benchmark Model classification report

The results in figure VI shows that the base author model achieved a total accuracy of 80%, which means it correctly classified 80% of the 2,250 images in the test set. The macro and weighted averages for precision, recall, and F1-score are all 0.80, indicating a reasonably balanced predictive performance across the three classes.

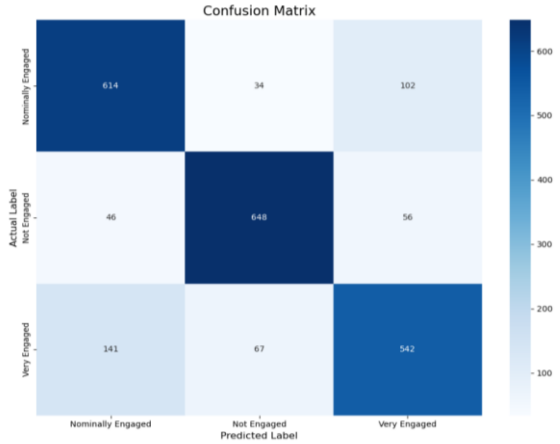

Figure VII Benchmark Confusion Matrix on Local Datasets

Why the MLP Outperformed the LSTM: Two factors explain the MLP's advantage. First, engagement in the 30-minute lectures changed slowly, so the temporal context provided by the LSTM offered little benefit over frame-wise classification. The LSTM's temporal memory smooths out rapid facial changes, which helps detect steady disengagement (identical recall of 0.86 for 'not engaged') but hurts recognition of brief, strong positive expressions that characterize high engagement (recall of 0.72 for 'very engaged' vs. 0.77 for the MLP). Second, the MLP's dropout regularization (p=0.3) prevented overfitting more effectively than the LSTM's more complex architecture, which has a higher capacity and therefore requires more data to generalize well. With only 6,300 training images, the LSTM is more prone to overfitting than the MLP.

For real-time applications where instructors want to recognize highly engaged students (e.g., to call on them or praise them) and quickly identify disengaged students, the MLP is preferable due to its higher recall for 'very engaged' (0.77 vs. 0.72) and higher precision for 'not engaged' (0.89 vs. 0.85). The lightweight MLP classifier runs in under 1 ms per frame on modest hardware (Intel Core i7, RTX 3060), making real-time feedback feasible even with basic computing resources. Moreover, the interpretable cues (specific emotions, drowsiness, phone use) address the transparency gap: instructors can understand precisely why a student was flagged.

Detailed Performance Comparison

Table IV presents a detailed performance comparison between the proposed DenseNet121+MLP model and the LSTM baseline on the balanced local test set. The MLP achieved an overall accuracy of 82% [79.5%, 84.2%], outperforming the LSTM at 80% [77.3%, 82.4%]. More importantly, the MLP showed consistent advantages across nearly all per-class metrics.

Discussion and Comparison

Table IV presents a detailed performance comparison between the proposed DenseNet121+MLP model and the LSTM baseline on the balanced local test set. The MLP achieved an overall accuracy of 82%, outperforming the LSTM (80%). More importantly, the MLP showed consistent advantages across nearly all per-class metrics.

Table IV: Detailed Performance Comparison of MLP vs. LSTM on Local Dataset

Model	Accuracy	Class	Precision	Recall	F1
MLP (Proposed)	82%	Not Engaged	0.89	0.86	0.88
		Nominally Engaged	0.81	0.83	0.82
		Very Engaged	0.79	0.77	0.77
LSTM (Baseline)	80%	Not Engaged	0.85	0.86	0.86
		Nominally Engaged	0.79	0.82	0.80
		Very Engaged	0.76	0.72	0.74

For the "not engaged" class, the MLP achieved higher precision (0.89 vs. 0.85) while matching recall (0.86). This means the MLP produced fewer false alarms (predicting disengagement when the student was actually engaged) without missing more true disengaged students. For instructors, this translates to more trustworthy alerts. For the "nominally engaged" class, the MLP again showed higher precision (0.81 vs. 0.79) and slightly higher recall (0.83 vs. 0.82), leading to a better F1 (0.82 vs. 0.80). The most notable difference was for the "very engaged" class: the MLP outperformed the LSTM in precision (0.79 vs. 0.76), recall (0.77 vs. 0.72), and F1 (0.77 vs. 0.74). The LSTM's lower recall for "very engaged" indicates that it missed many genuinely highly engaged students, likely because its temporal smoothing suppressed brief, strong positive expressions.

Evaluation on DAiSEE Benchmark

To test the generalizability of both models and to directly evaluate the hypothesis that class balance is more critical than architectural complexity, we evaluated both the proposed MLP and the LSTM baseline on the public DAiSEE dataset (Gupta et al., 2016). DAiSEE contains 9,068 video clips labeled with four engagement levels: very low (0), low (1), high (2), and very high (3). The dataset is severely imbalanced: the majority of samples belong to the "high" (56.2%) and "very high" (27.1%) classes, while "low" (9.3%) and "very low" (7.4%) classes are significantly under-represented.

Both models were evaluated on the official DAiSEE test split without any data augmentation or class rebalancing, to reflect real-world deployment conditions where class distributions may be skewed.

Table V: Performance on DAiSEE Benchmark (Four-Class Classification)

Model	Accuracy	Class	Precision	Recall	F1
MLP (Proposed)	61.2%	Very Low	0.18	0.12	0.14
		Low	0.29	0.21	0.24
		High	0.68	0.72	0.70
		Very High	0.62	0.58	0.60
LSTM (Baseline)	62.1%	Very Low	0.19	0.14	0.16
		Low	0.31	0.23	0.26
		High	0.69	0.73	0.71
		Very High	0.63	0.59	0.61

Both models achieved modest overall accuracy (MLP: 61.2%, LSTM: 62.1%), which is consistent with the state-of-the-art on DAiSEE (62–64%, Shiri et al., 2024; Mehta et al., 2022). However, the per-class metrics reveal a critical failure: both models performed poorly on the minority classes ("very low" and "low" engagement). The MLP achieved F1-scores of only 0.14 for "very low" and 0.24 for "low," while the LSTM achieved similarly low scores of 0.16 and 0.26. In contrast, both models performed well on the majority classes ("high" and "very high"), achieving F1-scores above 0.70.

This result has two important implications. First, it confirms that class imbalance is the primary barrier to accurate engagement detection, not architectural sophistication. Both a simple MLP and a more complex LSTM fail on minority classes when trained on imbalanced data. This supports our hypothesis that data quality and class balance are more critical than model complexity. Second, it validates our decision to collect and balance a local dataset. While our local dataset is smaller (n=38 participants, 9,000 images), the deliberate balancing of classes enabled the model to achieve high performance (82% accuracy, F1=0.88 for "not engaged") that is not possible on imbalanced public benchmarks.

Researchers and practitioners should prioritize collecting balanced, context-specific datasets over pursuing architectural innovations on imbalanced benchmarks. A simple CNN+MLP trained on balanced data can outperform sophisticated recurrent models on imbalanced data.

Discussion and Broader Implications

This study shows that a simple DenseNet121+MLP model, trained on local data, can detect engagement very well, with 82% accuracy. It can also make predictions in real-time, taking less than 1 millisecond per frame, and provide explanations for its predictions. These results are important for research on detecting engagement, especially in places with limited resources. They suggest that simple models can be effective and efficient, which is useful for real-world applications. The study's findings can help improve engagement detection in various settings, making it more accessible and practical.

What does it mean when we talk about local data being important? Let's look at an example. When we used a dataset from Kaduna, we got an accuracy of 82%. This is a lot better than the 62-64% accuracy that other studies got when they used a different dataset (DAiSEE). Now, it's not totally fair to compare these two datasets because they were labeled and defined in different ways. But the big difference in accuracy suggests that using local data that is balanced and relevant to the problem we're trying to solve is really important. This makes sense when we think about it; if we're trying to solve a problem in a specific place or community, it's better to use data from that place or community rather than data from somewhere else. This is something that people have been saying in the machine learning community for a while now: that data that is specific to a particular domain or problem is often more useful than bigger datasets that aren't as relevant.

Implication 2: Class Balance > Architectural Complexity. The DAiSEE results (Section 3.4) clearly demonstrate that when data is imbalanced, even sophisticated architectures fail on minority classes. The MLP and LSTM performed similarly poorly on "very

low" and "low" engagement classes, despite their architectural differences. This supports the argument that researchers should prioritize data collection, balancing, and annotation quality over chasing architectural improvements. Here's what it means for teachers to get real-time feedback that they can understand. The tool we're talking about is fast and can run on regular computers, not just the really powerful ones. This means teachers can get feedback right away, even if they're not using the latest and greatest technology. The feedback is also easy to understand, so teachers know exactly why a student is not paying attention. For example, they might see that a student is sleepy or using their phone, and they can do something about it. This is a big deal because it helps teachers know what's going on and how to help their students. In the past, teachers might just know that a student was not engaged, but they wouldn't know why. Now, they can see exactly what's going on and take action to help the student. This is especially important in places like Nigerian universities, where they might not have access to the latest technology. The tool is designed to be helpful in these kinds of settings, and it can make a big difference for teachers and students.

Implication 3: Practical Value for Under-Resourced Settings. In Nigerian tertiary institutions, where instructors often teach large classes (50–100+ students) with limited support staff, automated engagement alerts can help prioritize instructor attention. With a precision of 0.89 for the 'not engaged' class, an instructor receiving 10 alerts can trust approximately 9. In a typical 30-minute lecture, this translates to approximately 4–6 alerts, requiring about 2–3 minutes of instructor attention to investigate and intervene. This level of alert frequency is manageable and reduces the risk of alert fatigue.

Limitations of the Present Study

Several limitations of this study should be acknowledged.

First, the local dataset was collected from only 38 participants in Kaduna State, Nigeria, all attending synchronous lectures of 30 minutes duration. Whether the model generalizes to other regions of Nigeria, to secondary school students, to longer or asynchronous sessions, or to different cultural contexts remains unknown. The bootstrap confidence intervals (79.5–84.2% for accuracy) reflect uncertainty due to the small sample size.

Second, the system has a limitation, that it only uses video from a webcam. It doesn't take into account other important factors like audio, such as what students are asking or how they sound, or interaction logs, like what they're typing in a chat or how they're doing on quizzes. This means it might miss some important signs of how engaged students are. If it could combine all these different types of information, it might work even better, especially when the video isn't clear enough to make a good judgment. When people are watching and labeling how engaged someone is, there's always some room for error. Even though three different people did this job and mostly agreed with each other, it's still a pretty subjective task. There's just no way to avoid some ambiguity when you're trying to figure out what's going on in someone's head. And because of that, the labels that are supposed to be the "right" answers might not always perfectly match what's really going on with the person's engagement level.

Fourth, we didn't do a study to see if the model actually helps teachers teach better or if the real-time alerts make students learn more. Even though the technical side of things looks good, the real question is whether it makes a difference in how teachers teach and how well students do. The truth is, we need to know if this system really works in the classroom, and that's what matters most.

Fifth, the mild overfitting observed (95% training accuracy vs. 82% validation accuracy) suggests that the model may not generalize perfectly to new participants or settings. While dropout regularization mitigates this issue, a larger dataset would improve generalization.

Recommendations for Future Work

Based on the findings and limitations, we recommend several directions for future research.

First, to really get a good understanding of what's going on, we need to make our dataset bigger. This means getting more people involved; we're talking hundreds from all over Nigeria, and from different types of schools, like high schools, vocational schools, and universities. By doing this, we can make sure our findings are more widely applicable and see how different cultures express themselves. Having a bigger dataset, like over 500 people, would also allow us to do more detailed analyses and make sure our results are reliable. This would be a great step forward for future research.

Second, to really see if this system works, it's a good idea to try it out in a real classroom setting. This means using it in actual online courses and getting feedback from the instructors who are using it. We need to know if they find the alerts helpful and if the students are actually learning more because of it. This kind of test would give us the best proof that the system is useful in the real world. Just testing it in a lab isn't enough, we need to see how it works when it's being used by real people in a real classroom. By doing this, we can make sure that the system is really making a difference in how students learn.

Third, let's think about using multiple types of data together, like video, audio, and text. If we add audio features, such as how fast someone is talking, the tone of their voice, and if they're asking questions, and interaction features, like how active they are in a chat and how quickly they respond, we might be able to tell if someone is really engaged, even if they're not looking directly at something. For example, a student might be thinking deeply about a problem, but looking away, so just using video wouldn't catch that they're engaged. By using multiple types of data, we can get a more complete picture and reduce false alarms by checking cues across different types of data. This way, we can be sure that someone is really engaged or not.

Fourth, let's look into ways to protect privacy. We should think about using methods like federated learning or processing data right on the devices. This could help with the concerns people have about always being watched on video. If we process the data locally on the students' devices and only send alerts to the teachers, we can reduce the risks to their privacy.

Fifth, extend to longer sessions. Temporal modeling (e.g., LSTM, GRU) may become more beneficial with longer lectures (e.g., 90-minute sessions) where engagement states change more dynamically. Future work could compare MLP vs. LSTM on longer recordings to determine when temporal context becomes valuable.

Sixth, address class imbalance through synthetic data generation. While we balanced our local dataset through augmentation, future work could explore GAN-based synthetic data generation to create additional training samples for minority classes, further improving robustness.

Conclusion

This study developed and evaluated a DenseNet121+MLP model for predicting student engagement in online classes, using a locally collected dataset from Kaduna State, Nigeria, one of the first such efforts for a West African tertiary context. The model integrates four interpretable behavioral indicators: facial emotion recognition (seven categories), drowsiness detection, phone detection, and distraction detection. On a balanced local test set, the model achieved 82% accuracy and 0.88 F1-score for the most critical "not engaged" class, outperforming an LSTM baseline (80% accuracy). The system is lightweight enough for real-time inference (under 1 ms per frame) and its predictions are explainable via the specific cues that triggered them. The broader implication is that context-specific, locally collected data can yield substantially higher engagement detection accuracy than off-the-shelf models trained on distant populations. For instructors in under-resourced settings who cannot rely on expensive hardware or Western-trained models, our approach offers a practical path forward: collect modest amounts of local video, use a frozen pretrained CNN for feature extraction, and train a simple MLP classifier with dropout regularization. The resulting system can provide trustworthy, real-time, and interpretable engagement feedback, helping educators reclaim some of the instructional responsiveness that is lost when teaching moves from the physical classroom to the screen.

REFERENCES

Abedi, A., & Khan, S. S. (2021). Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network. *18th Conference on Robots and Vision (CRV)*, IEEE, 151–157.

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (0). Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 125.

Chandrasekar, Y. (2022). *Machine learning and eye-tracking framework to detect engagement in online learning* (Master's thesis). National College of Ireland.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Fredricks, J. A., Parr, A. K., Amemiya, J. L., Wang, M. T., & Brauer, S. (2019). What matters for urban adolescents' engagement and disengagement in school: A mixed-methods study. *Journal of Adolescent Research*, 34(5), 491–527.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Hossen, T., & Uddin, M. Z. (2023). Monitoring students' attention in online classes using machine learning approaches: A review. *Education and Information Technologies*, 1–26.

Huang, G., Liu, Z., Van der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Kainat, Ali, S., Iqbal, K. F., Avaz, Y., & Saiid, M. (2022). A review on different approaches for assessing student attentiveness in classroom using behavioural elements. *2022 2nd International Conference on Artificial Intelligence (ICAI)*, IEEE, 152–158.

Kansal, A. K., Gautam, J., Chintalapudi, N., Jain, S., & Battineni, G. (2021). Google trend analysis and paradigm shift of online education platforms during the COVID-19 pandemic. *Infectious Disease Reports*, 13(2), 418–428.

Maithri, M., Raghavendra, U., Gudigar, A., Samanth, J., Barua, P. D., Murugappan, M., & Acharya, U. R. (2022). Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine*, 215, 106646.

Mehta, N. K., Prasad, S. S., Saurav, S., Saini, R., & Singh, S. (2022). Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement. *Applied Intelligence*, 52(12), 13803–13823.

Ong, S. G. T., & Quek, G. C. L. (2023). Enhancing teacher–student interactions and student online engagement in an online learning environment. *Learning Environments Research*, 26(3), 681–707.

Orji, F. A., & Vassileva, J. (2022). Automatic modeling of student characteristics with interaction and physiological data using machine learning: A review. *Frontiers in Artificial Intelligence*, 5, 1015660.

Revadekar, A., Oak, S., Gadekar, A., & Bide, P. (2020). Gauging attention of students in an e-learning environment. *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, 1–6.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Sharma, P., Joshi, S., Gautam, S., Maharjan, S., Khanal, S. R., Reis, M. C., & de Jesus Filipe, V. M. (2022). Student engagement detection using emotion analysis, eye tracking and head movement with machine learning. *International Conference on Technology and Innovation in Learning, Teaching and Education*, 52–68.

Shiri, F. M., Ahmadi, E., Rezaee, M., & Perumal, T. (2024). Detection of student engagement in e-learning environments using EfficientNetV2-L together with RNN based models. *Journal on Artificial Intelligence*, 6(1), 85–103.

Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H., & Luo, A. (2021). Student class behavior dataset: A video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. *Neural Computing and Applications*, 33, 8335–8354.

Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., & Ali, L. (2023). Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data and Cognitive Computing*, 7(1), 48.

Vrochidis, A., Dimitriou, N., Krinidis, S., Panagiotidis, S., Parcharidis, S., & Tzovaras, D. (2024). A deep learning framework for monitoring audience engagement in online video events. *International Journal of Computational Intelligence Systems*, 17(1), 124.

Xu, Z., Yuan, H., & Liu, Q. (2020). Student performance prediction based on blended learning. *IEEE Transactions on Education*, 64(1), 66–73.