A Comprehensive Review of Vision-Based Student Engagement Monitoring Systems in Classroom Environments

Chaitanya Singh Bisht, Mehul Mahrishi, Sunil Dhankar, Shirish Nagar

Department of Computer Science Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur-302017 (India)

Email: csb.net.in@gmail.com, mehul@skit.ac.in, sunil@skit.ac.in, shirish.nagar@skit.ac.in

Received 04.03.2026; received in revised form 24.05.2026; accepted 24.05.2026

DOI: 10.47904/IJSKIT.16.1.2026.7-14

Abstract- Monitoring student engagement in classroom environments has become an active research area in computer vision, affective computing, and educational technology. Over the last decade, approaches have gradually shifted from specialised hardware-based tools to lightweight RGB-video pipelines that can operate in real time. This paper reviews twenty-two studies published between 2013 and 2026 covering automated classroom engagement monitoring, including motion and depth-based attention assessment, skeleton-based action recognition, facial expression analysis, multi-object tracking, YOLO-family behaviour detection, multimodal fusion of behaviour and emotion cues, dataset development, and deployment readiness. Each study is analysed in terms of modality, detection method, dataset, evaluation metric, deployment setting, and limitation. The reviewed literature is organised into six streams: early sensor-based systems, pose and skeleton-based methods, facial-expression approaches, behaviour detection models, multimodal fusion frameworks, and datasets or benchmarks. The comparative analysis shows that the field has made strong progress in real-time detection, but several issues remain unresolved. These include the absence of Indian classroom datasets, limited longitudinal per-student tracking, weak handling of back-row occlusion, insufficient on-premise deployment studies, and the lack of teacher-facing analytics. The review also highlights the need for privacy-preserving designs, consent-aware data collection, standardised benchmarks, and evaluation protocols that measure both technical accuracy and educational usefulness.

Keywords- Student engagement, classroom monitoring, literature review, behaviour detection, emotion recognition, YOLO, DeepSORT, on-premise deployment, privacy-preserving analytics.

1. INTRODUCTION

Student engagement is one of the most reliable indicators of academic performance, course completion, and learning quality. In large classrooms, however, a single instructor cannot continuously observe every student's posture, gaze direction, phone usage, note-taking behaviour, and facial expression. Signs of disengagement such as sleeping, repeated gaze aversion, or prolonged phone usage may therefore remain unnoticed until learning outcomes have already declined. This practical challenge has encouraged researchers to explore computer vision and machine learning as tools for monitoring engagement in a more systematic and scalable way.

Early systems in this domain used specialised hardware such as depth cameras, wide-angle cameras, and wearable sensors to capture student activity [1], [2]. These systems demonstrated the feasibility of automated attention assessment, but they were usually expensive, difficult to scale, and sensitive to camera placement. The later shift toward standard RGB video made it possible to use face detection, head-pose estimation, skeleton-based action recognition, and object detection as practical signals for classroom attention [3]-[6]. More recent work has used deep learning detectors such as YOLO-family models for real-time classroom behaviour recognition [11], [12], [15], along with facial expression recognition models for affective engagement analysis [7], [17]-[20].

Despite this progress, several gaps remain visible across the literature. Many studies focus on either behaviour or emotion, but not on a unified multimodal representation. Persistent student identity across a class session or across multiple sessions is rarely addressed. On-premise deployment, which is important for educational privacy and institutional control, is often missing from experimental studies. Indian-context emotion data is still limited, which may introduce demographic bias when models trained on general datasets are applied in Indian classrooms. Finally, teacher-facing analytics dashboards that translate model outputs into actionable classroom insights are still uncommon.

The aim of this review is to provide a comprehensive thematic analysis of twenty-two studies published between 2013 and 2026 on vision-based classroom engagement monitoring. The objectives are to categorise the literature into major research streams, compare the reviewed works across common dimensions, identify persistent technical and practical gaps, and suggest future directions for deployable and ethically responsible classroom monitoring systems.

The remainder of the paper is organised as follows. Section 2 describes the review methodology. Section 3 presents the thematic literature review. Section 4 provides the comparative summary. Section 5 discusses observed trends and limitations. Section 6 outlines research gaps and future directions. Section 7 concludes the paper.

2. REVIEW METHODOLOGY

This review follows a narrative and thematic review approach. The literature search was conducted using Google Scholar, IEEE Xplore, Scopus, MDPI journals, Springer databases, and relevant preprint repositories. The search terms included student engagement detection, classroom behaviour recognition, classroom monitoring computer vision, YOLO classroom, facial expression recognition in education, DeepSORT student tracking, multimodal engagement detection, and on-premise classroom monitoring system. The review covers publications from 2013 to January 2026.

The inclusion criteria were as follows: the study had to address automated monitoring of student engagement, attention, affect, or classroom behaviour; it had to use computer vision, image processing, deep learning, or a closely related visual analytics technique; and it had to be published in a peer-reviewed venue or made available as a recognised technical preprint. Studies that focused only on online learning or webcam-based engagement in video conferencing were excluded unless their methods had clear relevance to physical classroom monitoring.

Twenty-two studies were selected for detailed review. They were grouped into six thematic streams according to their main contribution: early sensor and depth-based systems, pose estimation and skeleton-based methods, facial expression recognition approaches, YOLO-based behaviour detection models, multimodal fusion frameworks, and datasets or benchmarks. Some studies contribute to more than one stream, but each is discussed under its primary contribution to maintain clarity.

3. THEMATIC LITERATURE REVIEW

3.1. Early Sensor and Depth-Based Systems

The earliest automated classroom monitoring systems relied on specialised hardware to capture student activity at a level of detail that standard cameras could not provide at the time. A wide-angle video-based system was used to assess classroom attention by measuring gross body movement and head orientation as attention proxies [1]. This work was important because it showed that automated attention estimation was possible in a lecture-hall environment. However, the approach required careful camera placement, offered only coarse behavioural cues, and did not provide individual student identification.

A later depth-based system used Microsoft Kinect sensors to combine skeletal information with facial landmarks for predicting student attention [2]. Depth sensing provided richer spatial information than RGB video alone, but the system depended on proprietary hardware and was evaluated on a relatively small classroom sample. As a result, its scalability and suitability for large, diverse classrooms remained uncertain.

These studies are foundational because they introduced the idea of automated attention assessment in physical classrooms. At the same time, they also reveal limitations that continue to influence current research: specialised hardware, small-scale validation, single-modality analysis, and limited deployment discussion.

3.2. Pose Estimation and Skeleton-Based Methods

As deep learning matured, researchers shifted from hardware-dependent depth sensing to software-based pose estimation on standard RGB video. A computer vision system combining person detection and pose analysis was proposed for student behaviour monitoring in classrooms [3]. This system could separate broad attentive and inattentive postures, which marked an important step toward practical RGB-based monitoring. However, the behaviour categories were still coarse and did not include emotion, identity tracking, or teacher-oriented reporting.

Skeleton pose estimation was later combined with person detection to recognise classroom behaviours such as writing, listening, and looking away [4]. Skeleton keypoints are useful because they reduce dependence on clothing appearance and lighting conditions. Nevertheless, keypoint detection becomes less reliable when students are partly hidden, especially in rear rows, making occlusion a significant challenge.

Visible engagement has also been studied from an educational psychology perspective using machine learning [5]. This work is valuable because it connects computational engagement measurement with learning theory and classroom observation. However, it is more conceptual than deployment-oriented and does not present a complete real-time monitoring system.

Lightweight real-time camera systems have been proposed to make attention monitoring feasible on resource-constrained devices [6]. Such systems demonstrate that classroom attention analysis can be performed without expensive equipment, but their simplicity often reduces behavioural granularity and does not address emotion recognition or persistent identity.

3.3. Facial Expression Recognition Approaches

A parallel research stream uses facial expressions as indicators of affective engagement. Real-time deep learning-based facial expression recognition has been applied to classroom settings to classify student expressions into standard emotion categories [7]. This approach can provide information that behaviour-only systems may miss. For example, a student may sit upright and look forward while still appearing bored or frustrated. However, emotion-only approaches do not capture behaviours such as phone usage, sleeping, or writing, and they generally do not include per-student tracking.

The effectiveness of facial expression recognition depends heavily on dataset quality and demographic diversity. FER2013 introduced a widely used facial expression benchmark with nearly 35,000 images across seven emotion categories [17]. Although it became a common baseline, it contains label noise and has limited demographic balance. RAF-DB improved annotation reliability and real-world variation through crowdsourced labels and deep locality-preserving learning [18]. AffectNet further expanded the scale of facial expression research by providing more than one million images with categorical emotion and valence-arousal annotations [19].

For Indian classrooms, demographic representation is especially important. IIITM Face was introduced for facial expression analysis in the Indian context and helps address bias caused by over-reliance on Western and East Asian datasets [20]. This is relevant because emotion models trained on non-representative datasets may confuse culturally or demographically specific neutral expressions with sadness or disengagement. Therefore, future classroom systems should use multi-source training data and evaluate performance separately across demographic groups.

3.4. YOLO-Based Behaviour Detection Models

The YOLO family of object detectors has become a dominant choice for real-time classroom behaviour detection because it provides a strong balance between inference speed and detection accuracy. A WAD-YOLOv8 method was proposed for detecting student behaviour in crowded classrooms, where students may overlap visually and occupy small regions in the frame [11]. The weighted attention distillation strategy improved detection in dense scenes, but the system focused on behaviour classes only and did not include emotion recognition, persistent identity, or a full classroom analytics application.

Improved YOLOv8s-based approaches have also been proposed for real-time classroom behaviour detection [12]. These methods focus on small-object detection and efficient inference, which are important for classroom videos where rear-row students may appear small. However, they still mainly address behaviour recognition as an isolated task.

Attention-enhanced YOLOv8 architectures have been applied to student classroom behaviour recognition [15]. Such models improve feature focus and category discrimination, but they generally do not combine behaviour with emotion, gaze, identity, longitudinal scoring, or teacher-facing reporting. A systematic review of classroom behaviour recognition also confirms that dataset limitations, deployment gaps, and inconsistent evaluation procedures remain major problems in the field [14].

The Ultralytics framework provides the training and deployment infrastructure for recent YOLO-family models, including YOLO26 variants used in classroom monitoring experiments [16]. Lightweight models such as YOLO26n are particularly relevant for on-premise or edge deployment because they reduce computational requirements while maintaining real-time feasibility.

3.5. Multimodal Fusion Frameworks

The most promising direction in classroom engagement monitoring is the integration of multiple signals into a single assessment framework. A real-time attention monitoring system combined behaviour detection, DeepSORT tracking, and emotion recognition [8]. This work demonstrated that combining behavioural and affective cues can make engagement estimation more reliable than using a single cue. However, it did not provide persistent identity management across sessions or a complete teacher-facing analytics interface.

Multimodal behaviour-emotion fusion models have been shown to outperform unimodal branches when measuring student engagement [9]. This supports the view that classroom engagement is not a single visual phenomenon. Behavioural cues show what a student is doing, while affective cues indicate how the student may be responding emotionally. A reliable system should combine both types of evidence rather than treating one as sufficient.

Multi-object tracking is an important component of multimodal systems. DeepSORT combines motion prediction with deep appearance features and has become a common tracking method in video analytics [21]. In classrooms, tracking helps connect detections across frames so that engagement can be measured over time rather than as isolated frame-level events. For student identification, ArcFace-based face recognition provides discriminative embeddings that can support persistent identity association when privacy and consent requirements are properly satisfied [22].

3.6. Datasets and Benchmarks

Datasets are central to progress in classroom engagement monitoring. The SCB-Dataset provides labelled examples for detecting student and teacher classroom behaviour and has been used by later behaviour recognition studies [10]. By including both student and teacher behaviours, it supports analysis of classroom interaction patterns rather than isolated student actions.

A video dataset for classroom group engagement recognition provides temporal information about collective attention at the group level [13]. This is useful because classroom engagement is often influenced by instructional events, group dynamics, and shared activities. However, group-level labels cannot replace individual-level longitudinal datasets, especially when the goal is to identify patterns of disengagement for specific students over time.

For emotion recognition, FER2013, RAF-DB, AffectNet, and IIITM Face provide complementary strengths in scale, annotation quality, and demographic coverage [17]-[20]. A robust classroom system should not depend on a single general-purpose emotion dataset. Instead, it should combine diverse sources and report performance using subject-level splits, demographic subgroup evaluation, and classroom-specific validation.

4. COMPARATIVE SUMMARY

Table 1 compares the twenty-two reviewed studies across nine dimensions: study, year, modality, core method, dataset, reported metric, real-time capability, deployment setting, and main limitation. The table is designed as a compact reference for researchers who want to position new work within the existing literature.

The comparison shows three broad patterns. First, studies before 2020 mainly relied on specialised hardware or single-modality analysis. Second, the 2021-2023 period saw more software-based and real-time methods. Third, studies from 2024 onwards increasingly use YOLO-based behaviour detection and multimodal concepts, but complete systems that combine behaviour, emotion, tracking, identity, privacy-preserving deployment, and teacher analytics remain rare.

Table 1: Comparative summary of reviewed studies on vision-based classroom engagement monitoring. Abbreviations: N/A = not applicable; V/A = valence/arousal; MOTA = multi-object tracking accuracy.

Study	Year	Modality	Core method	Dataset	Key metric	Real-time	Deployment	Main limitation
Raca and Dillenbourg [1]	2013	Movement and head orientation	Wide-angle video analysis	Custom lecture-hall data	Not reported	No	Lab only	Specialised camera setup; no identity tracking; coarse attention proxy
Zaletelj and Košir [2]	2017	Depth, face, and body posture	Kinect skeletal and facial landmark analysis	Small custom cohort	Not reported	Partial	Lab only	Kinect dependency; small-scale validation; limited modality coverage
Ngoc Anh et al. [3]	2019	RGB pose	Person detection with pose analysis	Custom classroom data	Not reported	Yes	Prototype	Binary attention states; no emotion recognition or tracking
Lin et al. [4]	2021	Skeleton	Skeleton pose estimation with person detection	Custom classroom data	Not reported	Yes	Prototype	Rear-row occlusion issues; no emotion module
Goldberg et al. [5]	2021	Visible engagement	Machine learning-based engagement assessment	Custom video data	Not reported	No	Conceptual	Strong theoretical grounding but not a deployed monitoring system
Renawi et al. [6]	2022	RGB attention	Lightweight camera-based models	Custom data	Not reported	Yes	Edge device	Reduced behaviour granularity; no emotion or identity management
Dukic and Sovic Krzic [7]	2022	Facial expression	Deep learning facial expression recognition	Custom and FER-based data	Not reported	Yes	Prototype	Emotion only; no behaviour fusion or tracking
Trabelsi et al. [8]	2023	Behaviour, emotion, and tracking	Detection, DeepSORT, and emotion recognition	Custom and FER-based data	Not reported	Yes	Prototype	No persistent student identity; no teacher-facing interface
Mahmood et al. [9]	2024	Behaviour and emotion	Fusion model comparison	Custom classroom video data	Fusion outperformed unimodal branches	No	Research	No tracking; no deployment study; no Indian-context data
Yang [10]	2024	Behaviour	Dataset contribution for student and teacher behaviour	SCB-Dataset	N/A (dataset)	N/A	N/A	Dataset only; no complete monitoring system
Han et al. [11]	2025	Behaviour	WAD-YOLOv8	Custom dense classroom data	Strong mAP	Yes	Research	Behaviour only; no emotion, tracking, or application layer
Sheng et al. [12]	2025	Behaviour	Improved YOLOv8s	Custom data	Competitive mAP	Yes	Research	Behaviour only; no multimodal fusion

Study	Year	Modality	Core method	Dataset	Key metric	Real-time	Deployment	Main limitation
Lu et al. [13]	2025	Group engagement	Video dataset contribution	Custom group dataset	N/A (dataset)	N/A	N/A	Dataset only; group-level labels rather than individual tracking
Liu et al. [14]	2025	Survey	Systematic review of computer vision methods	N/A	N/A	N/A	N/A	Review identifies gaps but does not implement a system
Zhang et al. [15]	2025	Behaviour	Attention-enhanced YOLOv8	Custom data	Strong accuracy	Yes	Research	Behaviour only; no emotion recognition or deployment layer
Ultralytics [16]	2026	Detection framework	YOLO26 architecture and training framework	N/A	N/A	Yes	Framework	Infrastructure framework rather than a classroom-specific system
Goodfellow et al. [17]	2015	Facial expression	FER2013 dataset	FER2013 (about 35K images)	N/A (dataset)	N/A	N/A	Label noise; limited demographic diversity
Li et al. [18]	2017	Facial expression	RAF-DB dataset and deep learning	RAF-DB	Improved over FER2013	N/A	N/A	Not classroom-specific
Mollahosseini et al. [19]	2019	Facial expression and V/A	AffectNet database	AffectNet (1M+ images)	N/A (dataset)	N/A	N/A	General-purpose dataset; not classroom-specific
Sharma et al. [20]	2023	Indian facial expression	IIITM Face database	IIITM Face	N/A (dataset)	N/A	N/A	Indian-context dataset but limited in scale
Wojke et al. [21]	2017	Multi-object tracking	DeepSORT algorithm	MOT benchmarks	Improved MOTA	Yes	Framework	General tracking method; not classroom-adapted by itself
Deng et al. [22]	2019	Face recognition	ArcFace loss and embeddings	LFW, CFP, and AgeDB	99.83% LFW	Yes	Framework	General face recognition method; not classroom-specific by itself

5. ANALYSIS AND DISCUSSION

5.1. Temporal Trends

The reviewed literature shows a clear evolution in classroom engagement monitoring. The earliest period, from 2013 to 2017, was dominated by hardware-based solutions such as wide-angle cameras and depth sensors [1], [2]. The middle period, from 2019 to 2022, moved toward software-based methods using RGB cameras, pose estimation, and lightweight deep learning [3]-[7]. From 2023 to 2026, research increasingly focused on YOLO-based real-time detection, multimodal fusion, and integrated pipelines [8]-[16]. This shift reflects the broader trend in computer vision, where specialised hardware is gradually being replaced by software-defined approaches that can run on standard cameras and edge devices.

5.2. Modality Analysis

Most reviewed studies still focus on a single modality. Behaviour detection is useful for identifying visible actions such as writing, phone usage, looking away, or sleeping, while facial expression recognition provides affective cues such as boredom, happiness, sadness, or frustration. Neither modality is sufficient by itself. A student may appear behaviourally attentive while emotionally disengaged, or may briefly look away while still being cognitively engaged. This is why multimodal fusion is important for practical engagement monitoring [8], [9].

5.3. Dataset Limitations

The dataset landscape contains both strengths and weaknesses. Emotion recognition benefits from established benchmarks such as FER2013, RAF-DB, and AffectNet [17]-[19], but these datasets were not designed specifically for classrooms and do not fully represent Indian students. IIITM Face helps address Indian-context representation [20], but larger and more diverse Indian classroom datasets are still needed. For behaviour detection, datasets such as SCB-Dataset [10] and several custom classroom datasets provide useful training material, but there is still no widely accepted benchmark comparable to COCO or ImageNet for classroom behaviour recognition.

Dataset standardisation is also necessary. Future datasets should define consistent behaviour categories, annotation rules, class imbalance handling, subject-level train-test splits, demographic metadata, camera positions, and classroom conditions. Without such standards, reported metrics across studies remain difficult to compare.

5.4. Deployment Gap

One of the strongest gaps in the literature is the limited discussion of deployment-ready systems. Many studies report detection metrics on recorded datasets but do not describe how the system would operate in an actual classroom. Practical deployment requires real-time inference, persistent storage, session management, secure authentication, teacher dashboards, failure handling, and hardware-aware optimisation. These engineering concerns are often

as important as model accuracy for real-world adoption.

On-premise deployment is especially relevant for educational institutions because student video, face images, and engagement records are sensitive data. Cloud-based processing may simplify computation, but it raises concerns about institutional control, privacy, and regulatory compliance. Future systems should therefore evaluate local processing, containerised services, encrypted storage, and restricted access as part of the research design.

5.5. Tracking and Identity

Multi-object tracking in classrooms has unique challenges. Students are often seated and move less than pedestrians in surveillance videos, which makes motion prediction easier. However, re-identification can be difficult when students are partially visible, appear similar from the back, or are occluded by other students. DeepSORT provides a useful tracking foundation [21], but classroom-specific tracking evaluation is still limited.

Persistent identity management across sessions is another underexplored area. Face recognition methods such as ArcFace can support student identification [22], but their use in education must be governed carefully. Identity linking should be used only when there is a legitimate educational purpose, informed consent, secure storage, and clear limits on access and retention.

5.6. Teacher-Facing Analytics

Most reviewed studies stop at technical evaluation and do not show how a teacher would actually use the output. Accuracy, mAP, AUC, and recall are useful research metrics, but teachers need interpretable session summaries, engagement trends, attendance-linked insights, and alerts that support timely intervention. A practical system should convert frame-level detections into understandable educational information rather than presenting raw computer-vision outputs.

Teacher-facing analytics should also avoid punitive surveillance. The goal should be to support learning and identify students who may need help, not to create unnecessary monitoring pressure. This requires careful interface design, explainable outputs, and institutional policies that define appropriate use.

6. RESEARCH GAPS AND FUTURE DIRECTIONS

6.1. Gap 1: Integrated Multimodal Systems

Most existing research studies address either behaviour recognition or emotion recognition. Future work should focus on integrated pipelines that combine behavioural, emotional, gaze-based, and temporal signals into reliable engagement scores. EMA-based scoring is one promising direction because it smooths short-term noise and reflects engagement over time, but it should be standardised and validated across different classroom settings.

6.2. Gap 2: Indian-Context Datasets

The lack of Indian classroom data remains a major limitation. IIITM Face provides a useful foundation for Indian facial expression analysis [20], but future datasets should include real classroom lighting, seating arrangements, camera angles, regional diversity, and age diversity. Subject-level splits should be mandatory to prevent leakage between training and testing. Dataset documentation should also describe consent procedures, demographic distribution, annotation rules, and known limitations.

6.3. Gap 3: On-Premise Deployment

Classroom monitoring systems should be designed for secure local deployment whenever possible. Future research should address model compression, quantisation, edge inference, containerised deployment, efficient database design, role-based access control, and local audit logging. A Docker-based microservice architecture with local analytics and optional local language-model support is a practical direction for institutions that do not want student data to leave campus.

6.4. Gap 4: Longitudinal Per-Student Tracking

Many current systems operate at the session level or frame level and do not maintain consistent student identity over time. Future systems should explore privacy-preserving tracking that can show how engagement patterns change across weeks or months. This would help teachers identify students who may be gradually disengaging and design timely academic interventions. However, longitudinal tracking must be accompanied by strict access control, retention limits, and institutional oversight.

6.5. Gap 5: Back-Row Occlusion

Back-row occlusion remains a difficult problem for single-camera systems. Students in rear rows may be partially hidden, small in the frame, or visually similar to nearby classmates. Future work should investigate elevated camera placement, multi-camera fusion, occlusion-aware detectors, synthetic data augmentation, and evaluation metrics that report front-row and rear-row performance separately.

6.6. Gap 6: Standardised Evaluation

The lack of standardised evaluation makes cross-study comparison difficult. Future benchmarks should define a common behaviour ontology, dataset splits, annotation guidelines, and baseline models. Technical metrics should include mAP@50, precision, recall, F1-score, AUC, Top-1 accuracy for emotion recognition, FPS, latency, IDF1 or MOTA

for tracking, and resource usage. Educational metrics should measure whether the system helps teachers understand classroom engagement and support students more effectively.

6.7. Gap 7: Ethics, Privacy, and Consent

Ethical design must be treated as a core research requirement rather than an optional deployment concern. Classroom monitoring involves students, and in many cases minors, so future systems should include informed consent, clear notice of recording, opt-out or alternative arrangements where feasible, and approval from institutional ethics committees. The purpose of monitoring should be limited to educational support, and the system should not be used for unnecessary surveillance or punitive profiling.

Privacy-preserving mechanisms should include data minimisation, local processing, encryption, strict role-based access, audit logs, short retention windows for raw video, and anonymised or pseudonymised analytics whenever individual identity is not required. Researchers should also report fairness analysis across demographic groups so that engagement scores do not disadvantage students based on appearance, lighting conditions, seating position, or cultural expression differences.

7. CONCLUSION

This review analysed twenty-two studies published between 2013 and 2026 on vision-based student engagement monitoring in classroom environments. The literature shows a clear movement from hardware-dependent and single-modality systems toward software-defined, real-time, and multimodal approaches that can operate on standard RGB video. YOLO-family detectors have become important for behaviour recognition, while facial expression datasets and multimodal fusion methods have improved affective engagement analysis.

Although the field has progressed, several gaps remain unresolved. These include limited multimodal integration, insufficient Indian-context data, weak longitudinal per-student tracking, back-row occlusion, limited on-premise deployment research, and the absence of standardised benchmarks. Future research should therefore prioritise dataset standardisation, subject-level splits, common behaviour taxonomies, benchmark evaluation protocols, and metrics that include detection accuracy, tracking consistency, real-time performance, demographic fairness, and educational usefulness.

The next generation of classroom engagement systems should be practical, privacy-preserving, and educationally meaningful. Such systems should combine behaviour, emotion, gaze, and temporal context; operate locally under institutional control; provide teacher-facing analytics; and support interventions that improve learning rather than merely recording student behaviour. If these requirements are addressed, vision-based classroom monitoring can become a useful support tool for teachers and students while respecting privacy, consent, and fairness.

REFERENCES

[1]. Raca, M., and Dillenbourg, P., "System for assessing classroom attention," Proceedings of the 3rd International Conference on Learning Analytics and Knowledge (LAK), Leuven, (2013), 271-275.

[2]. Zaletelj, J., and Košir, A., "Predicting students' attention in the classroom from Kinect facial and body features," EURASIP Journal on Image and Video Processing, (2017), vol. 2017, no. 1, 1-12.

[3]. Ngoc Anh, B., et al., "A computer-vision based application for student behavior monitoring in classroom," Applied Sciences, (2019), vol. 9, no. 22, 4729.

[4]. Lin, F.-C., Ngo, H.-H., Dow, C.-R., Lam, K.-H., and Le, H. L., "Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection," Sensors, (2021), vol. 21, no. 16, 5314.

[5]. Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., and Trautwein, U., "Attentive or not? Toward a machine learning approach to assessing students' visible engagement in learning," Educational Psychology Review, (2021), vol. 33.

[6]. Renawi, A., Alnajjar, F., Parambil, M., Trabelsi, Z., Gochoo, M., Khalid, S., and Mubin, O., "A simplified real-time camera-based attention monitoring system," Education and Information Technologies, (2022), vol. 27.

[7]. Dukic, D., and Sovic Krzic, A., "Real-time facial expression recognition using deep learning with classroom application," Electronics (MDPI), (2022).

[8]. Trabelsi, Z., Alnajjar, F., Parambil, M. M. A., Gochoo, M., and Ali, L., "Real-time attention monitoring system for classroom," Big Data and Cognitive Computing (MDPI), (2023), vol. 7, no. 1, 48.

[9]. Mahmood, N., Bhatti, S. M., Dawood, H., Pradhan, M. R., and Ahmad, H., "Measuring student engagement via behavior and emotion analysis in classroom videos," Algorithms (MDPI), (2024), vol. 17, no. 10, 458.

[10]. Yang, F., "SCB-Dataset: A dataset for detecting student and teacher classroom behavior," arXiv preprint arXiv:2304.02488, (2024).

[11]. Han, L., Ma, X., Dai, M., and Bai, L., "A WAD-YOLOv8-based method for classroom student behavior detection," Scientific Reports, (2025), vol. 15.

[12]. Sheng, X., Li, S., and Chan, S., "Real-time classroom student behavior detection based on improved YOLOv8s," Scientific Reports, (2025), vol. 15.

[13]. Lu, W., et al., "A video dataset for classroom group engagement recognition," Scientific Data, (2025), vol. 12.

[14]. Liu, Q., Jiang, X., and Jiang, R., "Classroom behavior recognition using computer vision: A systematic review," Sensors (MDPI), (2025), vol. 25, no. 2, 373.

[15]. Zhang, J., Guo, L., and Wang, X., "Student classroom behavior recognition based on YOLOv8 and attention mechanism," Information (MDPI), (2025), vol. 16, no. 11, 934.

[16]. Ultralytics, "YOLO26 documentation," Ultralytics, (2026).

[17]. Goodfellow, I. J., et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, (2015), vol. 64, 59-63.

[18]. Li, S., Deng, W., and Du, J., "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," Proc. CVPR, (2017).

[19]. Mollahosseini, A., Hasani, B., and Mahoor, M. H., "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Transactions on Affective Computing, (2019), vol. 10, no. 1, 18-31.
[20]. Sharma, N., Sharma, S., Mangla, M., Mohan, N., and Goyal, N., "IIITM Face: A database for facial expression analysis in Indian context," Multimedia Tools and Applications, Springer, (2023).
[21]. Wojke, N., Bewley, A., and Paulus, D., "Simple online and realtime tracking with a deep association metric," Proc. ICIP, (2017).
[22]. Deng, J., et al., "ArcFace: Additive angular margin loss for deep face recognition," Proc. CVPR, (2019).