

Continuous Engagement Detection in Smart Classrooms via Multicamera AI Systems in 5G-Enabled Educational Networks

Ana Paula S. Góis, Murilo Cruz Lopes, Mayara de Oliveira Paula, Douglas do Prado Pereira, Cristiani V. R. Guimarães, Francisco Martins Portelinha Junior, Felipe A. P. de Figueiredo, and Jose Marcos Camara Brito

Abstract—This paper presents an engagement detection system for smart classrooms leveraging a private 5G network and Multi-access Edge Computing (MEC). The architecture integrates IoT and computer vision to monitor student behavior through four strategically positioned cameras. By offloading video streams to the edge, the system executes classification of student actions, minimizing latency while maximizing processing responsiveness. Captured data is aggregated into a management dashboard that provides class-level engagement indicators based on predefined behavioral metrics. Results demonstrate that the 5G-enabled infrastructure ensures reliable data transmission and efficient remote management, providing a scalable proof-of-concept for intelligent educational environments.

Keywords—Student Engagement, Smart Classrooms, Computer Vision, 5G, Multicamera System, AI in Education

I. INTRODUCTION

The rapid evolution of digital communication and Artificial Intelligence (AI) drives the transition toward Education 4.0 [1], preparing students for the fourth industrial revolution through advanced robotics, Industrial Internet of Things (IIoT), and immersive environments. However, pre-5G networks suffer from data rate, latency, and reliability limitations, hindering the seamless delivery of sensory and auditory data essential for virtual learning [2].

In this context, 5G technology provides the necessary infrastructure for enabling smart classrooms [3]. Its core features, ultra-low latency (as low as 1ms) and massive connectivity, support innovative tools like Virtual Reality (VR) and Augmented Reality (AR) labs, which demand bandwidths exceeding 100 Mbps per user to prevent motion sickness caused by image delays [4]. Furthermore, integrating Multi-access Edge Computing (MEC) with 5G allows AI video processing to occur at the network edge. This proximity can reduce the system response times by up to 11.45% compared to traditional cloud-based approaches [3].

The integration of 5G also enhances real-time data collection through IoT devices and wearables, transforming the classroom into a network of context-aware agents [1]. This

infrastructure enables the synchronization of high-fidelity sensors, such as smart glasses and helmets equipped with accelerometers [5], surface electromyography (sEMG) armbands for gesture recognition [6], and EEG headbands for monitoring cognitive load [7]. While these wearables provide precise physiological data, including heart rate via PPG and electrodermal activity [8], their simultaneous use in crowded environments demands the ultra-low latency and massive connectivity that only 5G networks can provide [2].

However, a recurring challenge in the literature is the balance between data precision and user comfort. Although devices like smart pens, fitness bands, and haptic sensors offer deep insights into student engagement [5], [2], they can be intrusive or cost-prohibitive for large-scale deployment. Consequently, recent trends favor multimodal ecosystems where 5G-enabled AI acts as a non-invasive layer, utilizing smartphones as motion sensors [9] or high-definition cameras to complement wearable data. By offloading this massive sensory processing to the network edge, 5G transforms these isolated gadgets into a cohesive, real-time pedagogical management tool, overcoming the technical bottlenecks of previous mobile generations.

This work proposes a 5G-enabled Smart Classroom architecture that synchronizes four cameras using high-throughput private networks. By offloading streams to a Multi-access Edge Computing (MEC) framework, the system leverages 5G's low latency to run an integrated AI pipeline: RT-DETR for real-time person detection and MViT for behavioral classification. This 5G-AI convergence enables edge processing of complex neural networks, transforming raw data into an institutional dashboard for engagement insights. The resulting infrastructure automates attendance and access control through real-time pedagogical analysis, providing a scalable, anonymized framework for modern learning environments.

II. RELATED WORKS

Recent literature highlights the transformative role of 5G and AI in modernizing educational frameworks through enhanced connectivity and intelligent processing. Initially, research focused on AI and Machine Learning (ML) models for engagement detection. The CSFED [8] datasets established the foundation for quantifying student involvement by mapping affective states through static feature analysis. To address real-world classroom complexity, subsequent studies pivoted toward spatio-temporal tracking. Datasets such as SCSB [10] and CStudentAct [9] emphasized physical actions and teacher-

Ana Paula S. Góis, Murilo Cruz Lopes, Mayara de Oliveira Paula, Douglas do Prado Pereira, Cristiani V. R. Guimarães, Francisco Martins Portelinha Junior, are with the Inatel Competence Center (ICC), Santa Rita do Sapucaí, MG, Brazil. E-mail: ana.gois@inatel.br, murilo.lopes@inatel.br, mayara.paula@inatel.br, douglas.p@inatel.br, cristiani.vilela@inatel.br, franciscoportelinha@inatel.br. Felipe A. P. de Figueiredo is with the Wireless and Artificial Intelligence Laboratory (WAI Lab), Santa Rita do Sapucaí, MG, Brazil. José Marcos Câmara Brito are with the National Institute of Telecommunications (INATEL), Santa Rita do Sapucaí, MG, Brazil. This work was supported by RNP and MCTI.

student interactions, while [7] integrated emotional states to refine attention scoring.

The integration of 5G and Edge Computing has direct implications for pedagogical management. [3] proposed a teaching quality evaluation system that performs AI inference on local edge nodes rather than the cloud, reducing response latency by 11.45% and enabling immediate feedback. Complementing this approach, [11] introduced a 5G+AR model to deliver interactive situational teaching to rural areas, using Big Data to monitor student attentiveness in real-time.

III. METHODOLOGY

The proposed architecture follows a multi-layered framework for continuous behavioral analysis while preserving data integrity and student anonymity. As shown in Figure 1, the workflow is organized into three functional domains: the 5G, Computer Vision, and Dashboard layers. The Figure 1 also presents the prototype classroom at Inatel and the four installed cameras, with the fourth camera providing the displayed view. This workflow enables the transformation of raw video into educational insights across the three layers.

A. 5G Infrastructure (5G Layer)

The foundation of the system consists of cameras connected to a "5G-in-a-Box". This private 5G infrastructure is essential to handle the high bandwidth required by multiple simultaneous high-definition video streams. The 5G network ensures stable, low-latency transmission, which is critical in a classroom environment.

In conventional mobile networks, the functionalities of the Base Transceiver Station and the core are implemented on separate hardware platforms, resulting in a high-cost solution [12], [13].

The "5G-in-a-Box" solution integrates the 5G access node (gNodeB) and the 5G core (5GC) into a single compact unit. This solution is fully compatible with Release 15 Non-Standalone and Release 16 Standalone standards defined by the Third Generation Partnership Project (3GPP)[14]. This new paradigm enables the proliferation of numerous localized networks, in contrast to the current model dominated by a few global networks, and the creation of new applications [15]. Core network functions are virtualized, and the architecture of the equipment is based on Software-Defined Radio (SDR), enabling the entire protocol stack[16]. This approach simplifies updates, eliminating hardware modifications for evolution.

This streamlined infrastructure integrates key modules for 5G operation: a DC/DC Converter for power management; Baseband Processing for digital signals; a 3GPP Protocol Stack for standardized communication; and Software-Defined Radio (SDR) with RF Modules for radio transmission and reception. This modular design reduces hardware complexity while improving adaptability and cost-effectiveness, making it a robust solution for dynamic 5G private network environments.

Table I provides a summary of the operating specifications of the proposed solution, highlighting its key technical and functional features.

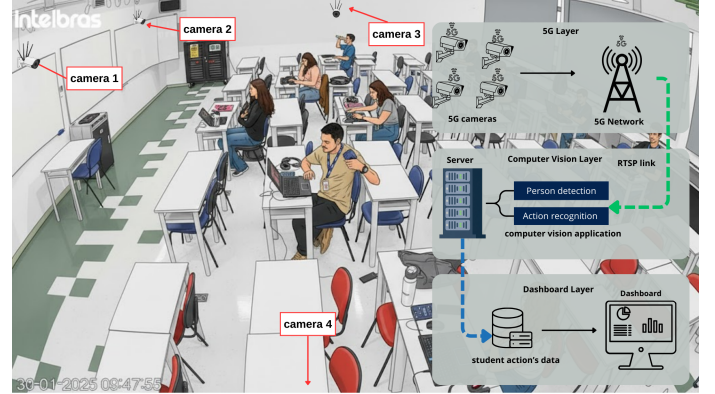


Fig. 1: Conceptual illustration of the multi-layered application architecture.

The 5G private network provides the infrastructure needed for cameras to connect and stream data to a local server.

B. Computer Vision Layer

1) *Dataset*: The dataset used in this study was constructed from primary footage recorded from the four cameras, see Figure 1, in real-world classroom environments over the course of two academic semesters in the Inatel Campus. A total of 468 hours of video were recorded, consisting of 60-minute. These recordings were selected, cut into 5-minute segments, and imported into CVAT (Computer Vision Annotation Tool), where manual annotations were performed to support two distinct tasks: student detection and action classification.

2) *Ethics and Data Privacy*: This study was approved by the Institutional Research Ethics Committee (Protocol FACULDADE DE CIÊNCIAS MÉDICAS DR. JOSÉ ANTÔNIO GARCIA COUTINHO - FACIMPA CAEE: 95261926.0.0000.5102). Before data collection, all participants provided informed consent authorizing the use of the recordings exclusively for scientific research and model development.

3) *Student detection*: The pre-trained RT-DETR model was fine-tuned to improve detection in crowded classrooms, with bounding boxes adjusted to include desks for context on student-object interactions. To reduce redundancy from near-static frames, training involved frame-skipping techniques. To ensure generalization and prevent data leakage, a date-wise (cross-session) partitioning was adopted, separating training and validation sets by different days/sessions. This ensures the model is evaluated under varying lighting and seating conditions.

4) *Student Tracking*: For the student tracking, this project employed ByteTrack, a state-of-the-art multi-object tracking algorithm, due to its strong performance reported on MOT17

TABLE I: Operating Specifications of the 5G Private Network.

Parameter	Description
Op. Band / Freq.	n78 (3.3 – 3.8 GHz)
Technology	5G NR
Bandwidth	100 MHz
Duplex Mode	TDD
Transceivers	MIMO 4T4R
Tx Power / Ant.	10 W
Supply Voltage	–48 VDC

and MOT20 benchmarks. The student tracking is important to retain the ID of the student during the entire class and also in the construction of the video clip to be applied on the video classification model [17].

5) *Multicamera Homography*: To recognize the same student across two different cameras, a homography-based strategy was employed. Homography is a projective transformation represented by a 3×3 matrix that maps points belonging to the same plane across different images [18]. In the classroom, the chairs were used as reference points, enabling the application of homography.

6) *Student Action Classification*: For the action classification task, a set of video classification neural networks was evaluated to identify the model that performed best on the training data. The evaluated models were *X3D*, *MViT_v2_s*, *MViT_v1_b*, *Swin3D*, and *SlowFast*. Hyperparameter optimization was performed to identify the best configuration for each model. The tested values included learning rates of 2×10^{-5} and 5×10^{-5} , *warmup* periods of 3 and 5 epochs, and *dropout rates* of 0.2 and 0.4. The *AdamW* optimizer was used together with the *Cross Entropy* loss function. Data augmentation techniques were also applied, including *RandomCrop*, *Horizontal Flip*, and *MixUp* with α values of 0.4 and 0.8. The dataset used in this stage was constructed using both frontal and lateral camera views. To define the action classes for the dataset, this study conducted a review of classroom behaviors based on recent literature [8], [9], [10]. Although additional behaviors were considered, a representative subset of actions was selected to provide a controlled evaluation while maintaining reliable model training. The selected classes and the total number of samples are detailed in Table II. To ensure a robust training process, the dataset was curated to maintain a uniform distribution, with each class comprising approximately 5,000 video clips. Train, validation, and test were separated in 70% train, 15% validation, and 15% test. Train and validation data were partitioned by calendar days in order to prevent data leakage, as surplus data from days assigned to validation or testing were excluded from the training phase.

TABLE II: Dataset video clip distribution per class.

Class	Train	Val	Test
On Phone	5,581	1,008	1,000
Paying Attention	5,632	1,084	1,093
Reading	3,801	1,024	1,038
Writing	4,903	1,006	1,027
Moving	4,784	1,107	1,025
On Computer	4,908	1,037	1,068
Total	29,615	6,182	6,257

C. Dashboard Layer

Following [8], detected actions are grouped into three observational categories. EL1 (Moving) represents non-engaged behavior, while EL3 (Paying Attention, Writing, and Reading) represents engaged behavior. Since On Phone and On Computer cannot reliably distinguish learning-related from unrelated activities, they are assigned to the neutral category

(EL2) and do not directly influence the engagement score. Aggregating these classifications produces the classroom Observational Engagement Profile.

IV. RESULTS AND DISCUSSION

A. 5G discussion

The proposed smart classroom relies on continuous real-time video transmission, requiring a communication infrastructure capable of providing reliable connectivity and communication performance suitable for real-time video transmission to support computer vision applications. Therefore, the private 5G network was validated before deploying the complete monitoring system.

Two commercial Samsung Galaxy S21 5G smartphones (SM-G991B) were used as User Equipments (UEs). The association between each UE and the 5G base station was verified through logs generated by the 3GPP protocol stack. These logs were used to monitor radio link quality indicators, including the Channel Quality Indicator (CQI), Modulation and Coding Scheme (MCS), and Signal-to-Noise Ratio (SNR), while network throughput was evaluated using *iperf*. Communication latency and packet loss were monitored through periodic ICMP (PING) packets.

A continuous 12-hour reliability test was performed to evaluate communication stability. During this period, no disconnection events were detected, and the communication link remained continuously available throughout the experiment. These results demonstrate that the deployed private 5G network provides the reliable communication required for continuous real-time video transmission, as shown in Figure 2.

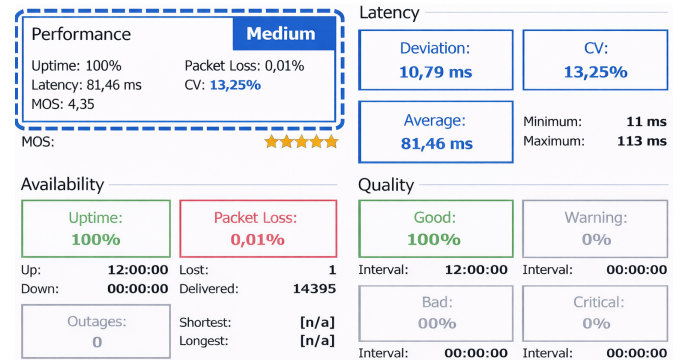


Fig. 2: Stability test (12-hour result).

After validating the communication infrastructure, the complete classroom scenario was evaluated using four cameras operating simultaneously over the private 5G network. The network sustained continuous video transmission without communication interruptions, demonstrating sufficient reliability and communication capacity for multi-camera acquisition in a smart classroom. The final deployment operated with four cameras configured at a resolution of 1920×1080 pixels using the H.264 codec, with an average bitrate of 615 kbps and an average frame rate of 28 FPS. Although the bandwidth required by the current deployment is modest (approximately 2.5 Mbps for four cameras), the proposed infrastructure was designed considering future smart classroom scenarios involving additional cameras, IoT devices, and mobile sensing

platforms sharing the same private communication network, thereby reducing susceptibility to the interference and spectrum contention characteristic of Wi-Fi networks operating in shared frequency bands.

B. Detection model

The performance comparison of different RT-DETRv2 architectures, considering different sampling steps, is presented in Table III. An insightful observation from these results is the relationship between model scale and data density.

While the smallest architecture, RT-DETRv2-S, achieved its peak performance using a Step 1 configuration (mAP@50 of 0.8670), all larger variants (M, L, and X) reached their highest metrics with Step 5. This suggests that for more complex architectures, a higher data density (Step 1) does not necessarily translate to better generalization, often leading to redundancy. The fact that larger models outperformed their Step 1 baselines when using Step 5 indicates that these models, for this classroom scenario, can achieve superior results with significantly less training data, optimizing the annotation and training pipeline.

Although RT-DETRv2-L (Step 5) achieved a slightly higher mAP@50 (0.8939), RT-DETRv2-S (Step 1) was selected for implementation. The 2.6% marginal gain of the L-variant does not justify its higher computational overhead in a multi-model environment. Thus, RT-DETRv2-S offers the optimal trade-off between accuracy and low-latency requirements.

TABLE III: RT-DETRv2 Results: mAP@50:95 / mAP@50

Metrics	mAP 50:95 / 50	mAP 50:95 / 50	50:95 / 50
Model	Step: One	Step: Five	Step: Ten
S	0.4181 / 0.8670	0.1857 / 0.5158	0.2447 / 0.6403
M*	0.3866 / 0.8338	0.4347 / 0.8744	0.4118 / 0.8745
L	0.4300 / 0.8584	0.4522 / 0.8939	0.4142 / 0.8561
M	0.3811 / 0.8184	0.3936 / 0.8420	0.3794 / 0.8338
X	0.4330 / 0.8725	0.4389 / 0.8820	0.3973 / 0.8488

* RT-DETRv2-M* indicates the R34-based medium model, while RT-DETRv2-M corresponds to the R50-m-based medium model.

C. Multicamera Homography

After applying the homography transformation to the side-camera image, the projected bounding boxes overlap with those in the frontal view, enabling the identification of the same student across both camera views. As shown in Figure 3(c), nearby projected bounding boxes indicate potential correspondences. The homography reduces the search space, and the Omni-Scale ReID model [19] uses a similarity threshold of 0.6 to determine whether the candidate detections belong to the same student. In the scenario where students are sitting in rows, the method could reach 78.57% of accuracy. In the scenario where students are sitting too close to each other, this accuracy can decrease to 60% due to this proximity and also due to the capacity of the detection model to separate the students, because two students in the same bounding box can confuse the REID model, that need one person per bounding box. Figure 3 c) where there are few students, and they are not close to each other, all the students detected in both cameras are related.

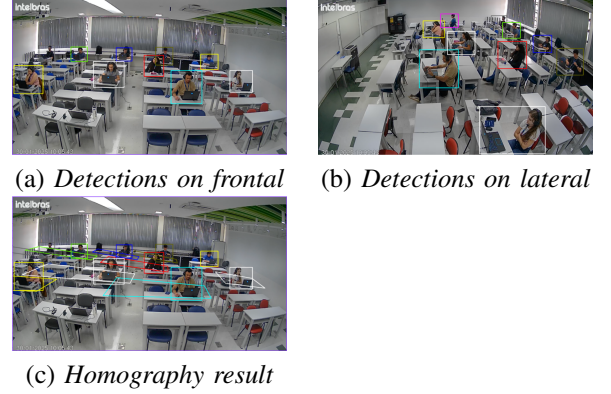


Fig. 3: Students with bounding boxes in frontal a) and Lateral b) cameras and lateral bounding boxes with the transformed perspective over the frontal camera c).

D. Classification model

Table IV presents the classification performance metrics for the test dataset. Accuracy, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC ROC) were analyzed for each model. The F1 metric was prioritized in the model selection and in the hyperparameter optimization process, due to its balanced assessment of precision and recall. The results demonstrate that the *MViT_v2_s* architecture exhibited superior performance across all evaluated metrics. The model achieved an accuracy of 74.78%, an F1-score of 74.43%, and an AUC ROC of 84.91%, representing a consistent improvement over all other tested architectures. The parameters that produced this results were a *Learning rate* of 5×10^{-5} , *weight decay* 0.05, *warmup* 0.5 and a *MixUp* of 0.4.

TABLE IV: Classification algorithm results with the test dataset and use of the best parameters found for the frontal database.

Algorithm	Accuracy (%)	F1-score (%)	AUC ROC (%)
<i>MViT_v1_b</i>	72.42	72.13	83.48
<i>MViT_v2_s</i>	74.78	74.43	84.91
<i>Swin3d_s</i>	71.73	71.26	83.08
<i>Slowfast_r101</i>	69.30	68.46	81.63
<i>X3D_M</i>	67.38	67.46	80.48



Fig. 4: Detection.



Fig. 5: Dashboard.

Fig 4 shows the application running, with the RT-DETR detecting the students, green bounding boxes, and the *MViT* model classifying the student behaviour, actions on the Bounding boxes.

For inference, a temporal window of 64 frames was used for each clip. Sampling is based on accumulating 64 consecutive bounding boxes per student generated by the detector and tracked using ByteTrack, ensuring sufficient temporal context for the model to capture behaviour patterns. Inference is performed using batch processing. Sequences from multiple

students are grouped into a single input tensor, optimizing GPU utilization and significantly reducing system response time. Inference was performed on an NVIDIA GeForce RTX 3070 Ti. Performance ranges from 15.41 FPS (40 students) to 24 FPS (15 students).

E. Application dashboard

The analytics layer features a Grafana-based dashboard, Fig. 5, that transforms 5G-acquired metadata into educational insights. The interface monitors real-time student count, dominant actions, such as Paying Attention, and engagement trends over time. This data-driven tool supports pedagogical decision-making by identifying peak engagement periods to optimize the academic calendar. By scheduling interactive content during high-energy intervals, the architecture facilitates a more dynamic and effective learning environment.

V. LIMITATIONS

During classroom recording tests, the cameras experienced intermittent and unsynchronized 5G connection losses. To address this issue, a 5G Customer Premises Equipment (CPE) was integrated into the system, providing wired Ethernet connectivity to each camera while preserving the benefits of the private 5G network. Extended experiments exceeding 100 hours confirmed the robustness and stability of the communication infrastructure.

Although the latency of the main processing modules was individually evaluated, this work did not measure the end-to-end latency from frame acquisition to dashboard visualization. Since the primary objective of this study was to validate the feasibility of deploying a private 5G network for intelligent classroom monitoring, a complete end-to-end latency analysis is left for future work.

Despite its effectiveness, the system has limitations due to the computational cost of deep learning models used for detection and behavior classification. The use of multiple models within the same application affects the overall performance of the computer vision system. As the number of students in the classroom increases, the application's frame rate decreases from 24 FPS to 15 FPS. This performance degradation is primarily attributed to the computational cost of the MViT model used for action classification.

Additionally, the system relies on homography to relate views across cameras and reduce ReID comparisons, but this requires selecting corresponding points between images, which limits scalability. Although methods like SuperGlue and SURF can assist in feature matching, their performance may vary with the environment, potentially affecting robustness in multi-classroom deployments.

VI. CONCLUSIONS

This research confirms that private 5G networks significantly enhance engineering education by enabling low-latency synchronization of multiple data streams. The proposed architecture supports a real-time AI pipeline, where the RT-DETR-S and MViT_v2_S models execute student detection and behavioral classification. Experimental results show

that transformer-based architectures, specifically MViT_v2_S, achieve superior performance across all scenarios. RT-DETR-S was implemented for its optimal balance between accuracy and inference stability. By integrating 5G with Edge Computing, the system processes multi-camera feeds into an automated dashboard for engagement and attendance tracking. This scalable solution demonstrates a robust framework for the digital transformation of classroom dynamics.

ACKNOWLEDGEMENTS

This work was partially funded by CNPq (Grant Nos. 444181/2024-5, 403612/2020-9, 311470/2021-1, and 403827/2021-3, 306199/2025-4); by Minas Gerais Research Foundation (FAPEMIG) (Grant Nos. PPE-00124-23, RED-00194-23, APQ-05305-23, and APQ-03162-24), by the Brasil 6G project (01245.020548/2021-07), supported by RNP and MCTI, by the projects XGM-AFCCT-2024-5-1-2 and XGM-AFCCT-2024-9-1-1 supported by xGMobile – EMBRAPII-Inatel Competence Center on 5G and 6G Networks, with financial resources from the PPI IoT/Manufatura 4.0 from MCTI grant number 052/2023, signed with EMBRAPII, and by FINEP (nº 1060/2 contract 01.25.0883.00).

REFERENCES

- [1] C. Koukaras et al., "AI-driven telecommunications for smart classrooms," *Telecom*, vol. 6, pp. 21, 2025.
- [2] B. Kizilkaya et al., "5G-enabled education 4.0: Enabling technologies, challenges, and solutions," *IEEE Access*, vol. 9, pp. 166962–166969, 2021.
- [3] F. Li and C. Wang, "Artificial intelligence and edge computing for teaching quality evaluation based on 5G," *J. Cloud Comp.*, vol. 12, p. 45, 2023.
- [4] L. Liujun, "The Impact of Integrating 5G Technology with University Campus Networks," *Sumerian J. Sci. Res.*, 2025.
- [5] D. Akila et al., "Research on recognition of students attention in offline classroom-based on deep learning," *Educ. Inf. Technol.*, vol. 29, pp. 6865–6893, 2024.
- [6] X. Sun, "5G joint AI technology in the innovation of university English education," *Wireless Comm. Mobile Comp.*, vol. 2021, p. 4892064, 2021.
- [7] Z. Trabelsi et al., "Real-Time Attention Monitoring System for Classroom," *Big Data Cogn. Comput.*, vol. 7, p. 48, Mar. 2023.
- [8] C. Pabba and P. Kumar, "An intelligent system for monitoring students' engagement," *Expert Systems*, vol. 39, no. 1, 2021.
- [9] P.-D. Nguyen et al., "A method for continuous student activity recognition from classroom videos," *Int. J. Mach. Learn. Cyber.*, vol. 16, pp. 7913–7937, 2025.
- [10] S. Zhang et al., "MSTA-SlowFast: A Student Behavior Detector for Classroom Environments," *Sensors*, vol. 23, p. 5205, 2023.
- [11] Z. Liu, "Based on 5G smart education to achieve education equality and wisdom," *Int. J. Educ. Hum.*, vol. 13, pp. 232–235, 2024.
- [12] D. Lohin et al., "The Disruptions of 5G on Data-Driven Technologies and Applications," *IEEE Trans. Knowl. Data Eng.*, vol. 32, pp. 1179–1198, 2020.
- [13] Y. Siriwardhana et al., "Performance Analysis of Local 5G Operator Architectures," *IEEE Internet Things J.*, vol. 7, pp. 11559–11575, 2020.
- [14] J. G. Souza, "Research and Tech. Development Project for an Integrated 5G Network: 5G-in-a-Box," Inatel, Utilization Plan, 2020.
- [15] L. Xu et al., "5G/5G-A Private Network: Construction, Operation and Applications," *Proc. IEEE TrustCom*, pp. 2234–2239, 2023.
- [16] A. G. Forte et al., "A Next-Generation Core Network Architecture for Mobile Networks," *Future Internet*, vol. 11, no. 7, 2019.
- [17] Y. Zhang et al., "ByteTrack: Multi-Object Tracking by Associating Every Detection Box," *arXiv:2110.06864*, 2021.
- [18] R. Szeliski, *Computer Vision: Algorithms and Applications*. London, U.K.: Springer, 2010.
- [19] K. Zhou et al., "Omni-scale feature learning for person re-identification," *arXiv:1905.00953*, 2019.