

DEEP LEARNING AND SENSING TECHNOLOGIES IN HIGHER EDUCATION: CLASSROOM BEHAVIOR RECOGNITION AND EVALUATION BY DA-DBN WITH LHHO OPTIMIZATION

Janaki Sivakumar^{1*}, Mehmet Aydin², Sudhanshu Maurya³, M. PrasannaBlessy⁴
Kirti Hemant Wanjale⁵, S. Agalya⁶

¹Associate Professor, Computer science and creative technology, Global college of engineering and technology, Muscat, Sultanate of Oman

²Senior Lecturer in Computer Science, Programme Leader for MSc Artificial Intelligence, University of the West of England Frenchay Campus, Bristol, UK.

³Department of Computer Science and Engineering, Manav Rachna International Institute of Research and Studies (Deemed to be University), Faridabad, India.

³Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India.

⁴Assistant Professor, Department of CSE (Cyber Security), Dr. Mahalingam College of Engineering & Technology, Pollachi-642003. India.

⁵Professor, Department of Computer Engineering, Vishwakarma Institute of Technology, Pune. India.

⁶Assistant Professor, Department of Information Technology, Dr. M.G.R. Educational and Research Institute. India.

¹<https://orcid.org/0000-0003-0412-1515>, ²<https://orcid.org/0000-0002-4890-5648>, ³<https://orcid.org/0000-0002-1999-1858>

⁴<https://orcid.org/0009-0005-8628-9590>, ⁵<https://orcid.org/0000-0003-4271-504X>, ⁶<https://orcid.org/0009-0007-6425-1748>

E-mail: *janakisivakumar2025@outlook.com, Mehmet.Aydin@uwe.ac.uk, dr.sm0302@gmail.com, blessypaul90@gmail.com, kirti.wanjale@vit.edu, agalyaselvakumar@gmail.com.

ARTICLE INFO

Article History

Received: November 25, 2025

Reviewed: December 24, 2025

Accepted: April 10, 2026

Published: June 30, 2026

Keywords:

Higher Education
Classroom Behavior
Student Behavior
Recognition
Evaluation

ABSTRACT

Student engagement plays a crucial role in facilitating comprehension in the teaching-learning process. Evaluating students' classroom behavior is essential for assessing the effectiveness of teaching. To objectively assess students' attentiveness in class, it is imperative to capture their emotions and cognitive feedback through expressions and behaviors. However, accurately extracting expression features from classroom videos poses a importantcontest. To discourse this, we suggest a hybrid optimization and deep learning technique used for higher education students. Our model involves designing a hybrid spatio-temporal ResNet with Laplacian Harris hawk optimization (LHHO) algorithm for feature extraction and optimization. It enables the extraction of contextual, semantic, and positional features while enhancing the recognition of occluded objects. Additionally, we employ a domain adaptive deep belief network (DA-DBN) for classroom behavior recognition and evaluation, improving the detection accuracy of occluded objects. To validate our approach, we evaluate it against existing classroom behavior recognition models using public datasets such as COCO and PASCAL VOC. The results demonstrate that our proposed algorithm accurately detects various student behaviors, outperforming state-of-the-art models. Specifically, LHHO+DA-DBN model achieves an average accuracy of 96.031% and a recall of 94.335%, surpassing the performance of compared models across different scenarios.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Classroom teaching is a pivotal aspect of modern education, particularly in disciplines like art education, where interaction, communication, and practical engagement among teachers and students are paramount [1]. Enhancing student contribution in schoolroom activities not only fosters complete artistic literacy but also nurtures their advanced and independent thinking skills [2]. However, outdated art schoolrooms often struggle with fully understanding each student's learning status, leading to ineffective teaching outcomes. Teachers face the challenging task of responding to separate student declarations and actions, monitoring overall class development, and adapting instructional objectives accordingly [3]. To succeed in this difficult role, teachers must visually perceive and procedure a vast quantity of learning-relevant information, a skill mentioned to as teacher perceiving in professional vision investigate. Numerous empirical studies in teacher research have focused on professional vision, particularly through video-based involvements where teachers observe classroom states and analyze noticed events, both in action and on action [4]. Expert-novice paradigm studies have highlighted significant alterations among pre-service and in-service teachers' observing of relevant classroom proceedings, with experienced teachers exhibiting a greater focus on scholars and key learning-related features of instruction [5]. Recent research has leveraged advanced eye-tracking methods to examine teacher noticing processes, revealing insights into the distribution of teachers' visual attention during classroom instruction [6].

For instance, experienced teachers tend to allocate attention more evenly among students and spend less time on task-redundant areas in the classroom compared to pre-service teachers. In art education settings, it's critical for educators to painstakingly screen understudies' exhibition and learning progress to speedily recognize any issues and deal fitting heading and backing [7]. Then, dissecting understudy conduct turns out to be especially fundamental in such homeroom conditions. By breaking down understudy conduct, instructors can acquire experiences into showing adequacy and give designated help where required [8] [9]. However, the diverse teaching scenarios, varying lighting, and wide range of student behaviors may make automatic detection of student learning behavior difficult in art teaching classrooms [10]. In current years, obsolete machine learning (ML) and deep learning (DL) approaches have shown superior show in perceiving understudy execution [11-14]. Traditional ML methods rely on feature extraction to generate behavior features for recognition, but they are limited by their representational abilities. DL-based student behavior recognition systems automatically learn deep features from large training datasets. These systems utilize DL models and parameter selection to achieve effective detection results. In the field of computer vision, there is a growing trend towards learning action representation using DL to enhance investigate on behavior gratitude in shadowing videos within schoolroom settings [15].

Behavior recognition aims to allocate a solitary class label to an complete cinematic, while spatio temporal behavior recognition involves labeling each person's achievement in each frame by deeply learning spatio temporal material from consecutive surrounds, which poses a greater challenge. Recent advancements have focused on relational reasoning-based behavior recognition methods to perfect the relationships among objects using boundary boxes in frames, as objects interrelate with each other [16]. Though, existing procedures often rely on appearance features extracted from individuals and matters detected by pretrained object detectors with fixed categories, potentially overlooking small objects connected to the actor. Additionally, in cluttered classroom environments, relational reasoning may prioritize salient regions in images, neglecting secondary human body portions crucial for human-to-human communication [17], [18]. Therefore, accurate detection of supporting human body parts and objects amid disorderly backgrounds is essential for interactive action recognition. Despite advancements, challenges persist in student behavior recognition [19]. Limited availability of datasets on students' learning behavior hampers research progress, with most studies focusing only on specific behaviors like rising hands, sleeping, and yawning, which are sovereign performances [20]. Though, the absence of interactive behaviors negatively impacts the assessment of scholars' learning excellence. To address these challenges, behavior recognition models focusing on accurate detection of interactive actions are proposed.

Our contributions. We propose a hybrid optimization and DL technique used for higher education students to address the problem of extracting features from classroom videos while it is posed. The major charities of planned work are given as follows.

1. To enables the extraction of contextual, semantic, and positional features while enhancing the recognition of occluded objects a hybrid spatio-temporal ResNet with Laplacian Harris hawk optimization (LHHO) algorithm for feature extraction and optimization.
2. To advance the detection accurateness of obstructed objects a domain adaptive deep belief network (DA-DBN) for classroom behavior recognition and evaluation.

The rest of this paper is prepared as follows: Section 2 reviewed the various models for classroom behavior recognition and evaluation using DL models using deep learning. The problem description and the detailed working process of proposed deep learning and sensing technologies in higher education is explained in Section 3. Section 4 examines the effectiveness of the proposed and existing classroom models using results and comparative analysis. Finally, the paper completes in Section 5.

II. RELATED WORKS

II.1 CLASSROOM BEHAVIOR RECOGNITION AND EVALUATION USING DL MODELS.

This section provides a brief overview of recent research on classroom behavior recognition and evaluation using DL models. According to [21] have developed a quick visually impaired remaking model with linear canonical transform (LCT) regularization for recreation of feeble infrared otherworldly signs. Using LCD conversion with high resolution and a weak IR spectral signal, they examined their primary distinction. At low resolution, bright IR spectra reveal a broad distribution of image data.

An infrared spectral reconstruction technique was developed in response to these findings to limit the wide range of IR spectra observed at L0-Nord. For human behavior monitoring, the experimental results demonstrate that the LCD-based IR spectrum reconstruction method can preserve IR spectral range structures and eliminate Poisson noise. For authors [22] have proposed an infrared look acknowledgment technique for understanding nonverbal ways of behaving in the study hall through multi-name conveyance learning. In particular, we originally determined the component closeness of seven fundamental looks to portray the connection between contiguous articulation pictures. Then, comparable qualities are fitted utilizing a Cauchy conveyance capability. Moreover, we build another profound organization utilizing Cauchy distribution-based label learning (CDLLNet) rather than the standard single articulation name. Utilizing these amended names, the infrared facial covering learns the adjoining demeanor marks and its unique appearance marks. A few subjective and quantitative experimental outcomes affirm that the CDLLNet network accomplishes vigorous outcomes and performs surprisingly well. In [23] have presented an extensive informational collection that can be utilized to recognize, analyze, and contextualize understudy conduct in the homeroom. An organized informational index comprises of a revelation section, a validation part, and a header part. The discovery region incorporates a transitory recognition information block with 4542 examples and an activity identification information block with 3343 examples, while the location region contains 4276 examples and the header region contains 4296 examples. In genuine homerooms, understudy conduct is programmed, making information assortment delegate and reasonable.

Then analyze specific aspects of the classroom situation and technical issues for each module, which are tested through tests. According to [24] have presented an extensive informational collection that can be utilized to recognize, analyze, and contextualize understudy conduct in the homeroom. An organized informational index comprises of a revelation section, a validation part, and a header part. The discovery region incorporates a transitory recognition information block with 4542 examples and an activity identification information block with 3343 examples, while the location region contains 4276 examples and the header region contains 4296 examples. In genuine homerooms, understudy conduct is robotized and information assortment is delegate and sensible. Then, for each experimented-on module, look at specific aspects of the classroom and technical issues. According to [25] have proposed a system for recognizing student behavior that uses personality detection and an assessment of skeletal posture utilized to exhibit the proposed framework. To reduce the number of false connections in bone data, a method of error correction based on techniques for person detection and pose estimation is proposed. Highlight extraction was performed to produce include vectors addressing the human figure. Qualities incorporate ordinary joint arrangement, joint dividing, and bone points. Social grouping was finished to decide the understudy's way of behaving. A profound brain network was worked to characterize the exercises and the proposed framework had the option to decide the quantity of understudies in the class. The framework accomplished 15.15% higher normal accuracy and 12.15% higher normal review than the skeleton-based conspire. For authors [26] have introduced to achieve an enhanced YOLO-v4 network, cross-local network embedded connection (EC) components are configured and embedded at the end of the YOLO-v4 network.

The added substance misfortune capability comprises of two sections: RepBox and RepGT losses. RepGT utilizes a misfortune capability to work out the misfortune values between the objective forecast box and the encompassing ground truth boxes to lessen bogus positive location results; RepBox utilizes the misfortune capability to work out the misfortune esteem between the objective expectation field and other close by target forecast fields to diminish misleading negative identification results. Preparing and testing are performed on homeroom conduct datasets of educators and understudies, individually. The outcomes show that the typical exactness for distinguishing the study hall conduct of various targets is over 90%, which demonstrates its viability. According to [27] have proposed a 3D-CNN is chosen as an organization model for homeroom understudy conduct acknowledgment, which has high precision and unwavering quality of multi-source data and further develops restriction of homeroom understudies. The spatio-temporal domain of 3D convolution extends 2D convolution for improved video analysis of human behavior. In classroom teaching situations, a deep spatial residual CNN is used to learn the spatial characteristics of each student's behavior and recognize classroom behaviors in real time. Different investigations are directed utilizing class conduct informational collections, and the experimental results show that the model performs well in identifying class behavior. In [28] have proposed Density, occlusion, and multimodal displays are just a few of the issues that the Student Classroom Behavior Framework addresses. They utilized a progression of imaginative techniques and joined the Res2Net block with the first C2f block in YOLOv8 to make the C2f_Res2block block.

The module productively handles multi-scale scenes and works on model precision. Multi-scale center and multi-head concentrate further work on model execution. According to [29] have proposed a technique for utilizing images of classroom time series to determine the active behavior of students. Students' active behavior is detected using Advanced asynchronous interaction aggregation (AIA). By adding the MsTA and MSCSA modules to the fast track and slow track, respectively, the enhanced AIA network backbone improves the accuracy of identifying student engagement behaviors in the slow fast track. A smooth focal loss function is used to enhance the contrast between segments in the pupil action database. The student activity behavior data set demonstrates that this method outperforms the original AIA network in terms of detection performance when it comes to identifying a variety of classroom student activity behaviors. According to [30] have proposed a data augmentation strategy for enhancing behavioral characteristics in order to identify distinctive student behaviors. The process begins with the collection of a narrow set of data designed to understand student learning behavior and expands it significantly using data augmentation techniques. The compositional abilities of extended-efficient layer aggregation network (E-Spirit) can be utilized to productively remove different learning conduct highlights. Outstandingly, the channel-wise attention module (CPAM) coordinates the central instrument into the component location organization. This works on the organization's capacity to recognize key highlights related with understudies' learning ways of behaving, consequently expanding the exactness of component acknowledgment. The finish of the excursion includes the characterization of highlights removed utilizing a double methodology: a feature pyramid network (FPN) and a way conglomeration organization (Skillet). Observational proof plainly exhibits the force of the technique acquiring 96.7% Guide.

II.2 RESEARCH GAP

In the province of classroom behavior gratitude and evaluation, numerous intricate challenges emerge, complicating the precisecalculatation of student attentiveness and appointment. The students' expressions and behaviors exhibit vast variability, encompassing facial expressions, body language, and verbal cues, which poses a challenge in their interpretation due to individual differences and cultural nuances. The presence of occlusions and environmental factors within classroom settings, such as students obstructing each other's view or varying lighting conditions, further complicates the recognition of these expressions and behaviors. The inherently subjective nature of evaluating attentiveness and engagement introduces another layer of complexity, as different educators may have divergent criteria for defining attentive behavior, leading to inconsistencies in assessment. Temporal dynamics also play a crucial role, with classroom interactions evolving over time, necessitating an understanding of how factors like lesson content and teaching style influence student behavior. Furthermore, privacy and ethical considerations surrounding the deployment of video-based monitoring systems in classrooms raise important concerns that must be carefully addressed. Data annotation and labeling present another challenge, as manually annotating classroom behavior data is resource-intensive and prone to biases. Moreover, ensuring the generalizability of behavior recognition models across diverse educational settings remains a significant hurdle, as nuances in student demographics and classroom dynamics can impact model performance.

Finally, effectively integrating behavioral insights into educational interventions and instructional design strategies poses yet another complex challenge in achieving the overarching goal of improving teaching practices and student outcomes. Moreover, the key challenges is the extraction of expression features from classroom videos, which is inherently complex due to various factors such as occlusions and varying environmental conditions. To discourse this contest, we suggest a original hybrid optimization and deep learning approach tailored for higher education students. Our proposed model comprises two main components: a hybrid spatio-temporal ResNet enhanced with the Laplacian Harris hawk optimization (LHHO) algorithm for feature extraction and optimization, and a domain adaptive deep belief network (DA-DBN) to further enhance the recognition of expression features. By adapting the feature representations learned from one domain (e.g., one classroom setting) to another domain (e.g., a different classroom setting), the DA-DBN ensures robust performance across diverse educational environments. Our proposed method represents a significant contribution to the field by offering a comprehensive solution for extracting expression features from classroom videos and assessing students' attentiveness in higher education settings. The integration of hybrid optimization techniques with deep learning architectures enhances the model's effectiveness and robustness, paving the way for more accurate and reliable evaluations of student engagement and learning outcomes.

III. PROPOSED METHODOLOGY

In the field of education, student participation is critical to improving understanding. When students actively engage with the information, ask questions, and interact with it, the teaching-learning process flourishes. Their participation creates a lively classroom atmosphere in addition to fostering understanding [31]. But evaluating this involvement calls for more than just watching; it also calls for a closer examination of the actions, facial expressions, and cognitive reactions of the students. Historically, educators have depended on nonverbal indications like engagement levels, body language, and eye contact. However, the development of technology has created new opportunities for evaluation.

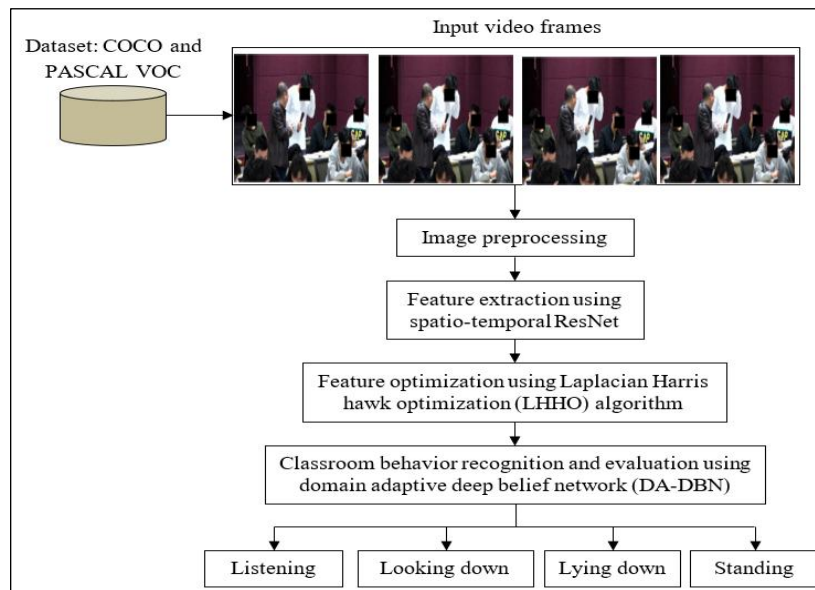


Figure 1: System design of proposed LHHO+DA-DBNmodel for classroom behavior recognition and evaluation.

Source: Authors, (2026).

Although classroom recordings provide an abundance of information, it is still quite difficult to extract useful expression elements from them. Emotions, cognitive input, and minute details must be extracted from video data using sophisticated algorithms and reliable techniques. Accuracy, privacy, and scalability must all be balanced as educators endeavor to understand the complex web of student interaction in the digital age. We suggest a hybrid optimization and deep learning method for college students as a solution to this problem. For feature extraction and optimization, our model builds a hybrid spatiotemporal ResNet using the Laplacian Harris Hawk Optimization (LHHO) algorithm. It improves the detection of obstructed objects and allows the extraction of contextual, semantic, and positional details. Furthermore, to enhance the detection accuracy of obscured objects, we utilize a domain adaptive deep belief network (DA-DBN) for classroom behavior recognition and evaluation. In the Figure 1 the proposed LHHO+DA-DBN model is given. For the analysis the video of a class is considered. By apply LHHO+DA-DBN method the behavior and features of the students in the class are identified. This method helps the academic authorities to make their facility more compactable for the students.

III.1 FEATURE EXTRACTION USING SPATIO-TEMPORAL RESNET

In the context of laboratory behavior gratitude and evaluation, feature abstraction mentions to the procedure of capturing relevant info from raw input data, such as video frames, to represent key characteristics or patterns of behavior. A spatio-temporal ResNet is a kind of deep neural network construction specifically calculated to process both spatial and temporal information. ResNet, short for Residual Network, is a popular type of convolutional neural network (CNN) building normally used in image recognition errands. In the case of spatio-temporal ResNet, the architecture is extended to handle video data, allowing it to imprisonment both longitudinal features and progressive topographies. The working process of feature extraction using spatio-temporal ResNet is given as follows.

- The input to the spatio-temporal ResNet is a sequence of video frames captured from a classroom setting. Each frame represents a snapshot of the classroom at a specific time.
- Before feeding the video frames into the network, preprocessing steps may be applied, such as resizing the frames to a unchanging size, standardizing pixel standards, and possibly enhancing the data to increase the assortment of the exercise dataset.
- Choose an appropriate spatio-temporal ResNet architecture based on the complexity of the task and available computational resources.
- Pass each frame through the spatio-temporal ResNet architecture to extract features. The network processes the frames in a hierarchical manner, with lower coatings apprehending low-level topographies and higher layers apprehending more intellectual topographies relevant to behavior recognition.
- Since behavior recognition involves analyzing temporal patterns over a sequence of frames, the features extracted from individual frames need to be aggregated over time. This can be done using methods such as chronological pooling or recurrent neural networks (RNNs) to capture temporal dependencies.
- After temporal processing, the extracted features are represented as high-dimensional feature vectors. These feature vectors encode information about the contextual, semantic, and positional aspects of the observed behaviors.
- In some cases, the dimensionality of the feature trajectories may be too high for downstream tasks such as classification. Dimensionality reduction technique is applied to reduce the dimensionality of the feature space while preservative vital material.
- The final output of feature extraction process is set of feature vectors, each corresponding to a specific segment of the input video sequence. These feature vectors can then be used as input to classifier for behavior recognition and evaluation.

Contextual features refer to information about the surroundings or environment in which the behavior occurs, providing additional context for understanding the behavior. Semantic features capture the meaning or semantics of the behavior, such as identifying specific actions or interactions. Positional features describe the spatial location or arrangement of objects or individuals within the scene. To extract these features using a spatio-temporal ResNet, the network is qualified on a large dataset of video sequences glossed with labels corresponding to different behaviors. During training, the network learns to automatically extract features from both individual frames and sequences of frames. The layers of the network are designed to capture hierarchical representations of the input data, gradually transforming raw pixel values into higher-level features that are more meaningful for behavior recognition. The final production of the system is set of feature vectors that represent the contextual, semantic, and positional aspects of the observed behaviors.

III.2 FEATURE OPTIMIZATION USING LAPLACIAN HARRIS HAWK OPTIMIZATION (LHHO)

Highlight enhancement is the method involved with choosing the most important and instructive elements from a given informational index to work on the presentation of AI models. A lot of real-world applications use a lot of datasets with a lot of features, some of which could be useless, redundant, or noisy. Include streamlining intends to distinguish and hold just the most discriminative elements, consequently lessening the size of the informational index and working on the exactness, interpretability, and execution of model expectations. Laplacian Harris Hawk Optimization (LHHO) is a metaheuristic optimization procedure stimulated by the shooting performance of Harris's hawks. It combines principles from the Harris hawk's collaborative hunting strategies with mathematical optimization techniques to efficiently search for optimal solutions to feature optimization problems. During the optimization process, LHHO employs a combination of exploration and exploitation strategies to iteratively update the candidate feature subsets. In the exploration phase, the algorithm encourages diversity by exploring different combinations of features across the search space.

This is achieved through random exploration and movement of candidate feature subsets, simulating the scouting behavior of Harris's hawks as they search for prey in varied terrain. LHHO algorithm maximizes the production of biodiesel by optimizing three independent variables. The objective function is depicted as follows.

$$\max d(z_s, z_n, z_v) = 37.45 + 0.42z_s + 4.37z_n - 0.16z_s^2 + 24.26z_v - 1.53z_v^2 + 0.001z_{sn} - 0.24z_s z_v - 0.44z_n z_v, \quad (1)$$

where $55 \leq z_s \leq 65$, $6 \leq z_n \leq 15$, and $0.6 \leq z_v \leq 2.0$.

Harris hawks typically have farsighted eyes, which enable them to spot and capture their prey. However, it is extremely difficult to easily identify prey. After an attack on the prey, hawks hold back, observe, and monitor for this reason. In this particular algorithm, candidate solutions are Harris hawks, and the optimal solution is taken into consideration at each step that is close to the ideal position. The exploration phase's mathematical representation is shown as

$$\begin{aligned} X_{rand}(o) - e_1, X(r+1) &= |X_{rand}(o) - 2t_2 Z(o)|, g \geq 0.5, \\ & (X(o) - X(o)), \\ t_3(lb + t_4(ub - lb)), & g \leq 0.5 \end{aligned} \quad (2)$$

where Z_{rabbit} is the position of the rabbit, X_{r+1} is the hawk's position, and p is the next iteration. The current location of the hawks is X_o , and the random numbers e_1 , e_2 , e_3 , e_4 , and h are between 0 and 1. The typical hawk's position is described as

$$X_b(o) = \frac{1}{A} \sum_{o=1}^A X_u(o), \quad (3)$$

where the position of each hawk in repetition p is x_u , and A denotes the total quantity of hawks. Transition from exploration to exploitation the exploratory phase can transition into the exploitative phase based on how the prey's energy escapes, allowing for a variety of exploitative behaviors. During the behavior of fleeing, the prey's energy decreases significantly.

$$\epsilon = 2 \epsilon_o \left(1 - \frac{o}{r}\right), \quad (4)$$

where ϵ signifies the absconding energy of the target, O denotes the initial energy state, and r denotes the supreme number of restatements. Laplacian numbers take the place of variant random numbers in LHHO. The reported Laplacian random number generator is used in recent methods by applying the matrix output of the function, which has a unkind value of 0 and normal nonconformity of 1.

$$\begin{aligned} E(A_{s,n}^f) &= [\text{cumsum}(2 \times \text{Lap}(r_1) - 1) \\ & \text{cumsum}(2 \times \text{Lap}(r_2) - 1), \dots, \\ & \text{cumsum}(2 \times \text{Lap}(r_{max}) - 1)] \end{aligned} \quad (5)$$

The conventional Laplace probability distribution (LPD), denoted by $\text{Lap}(S, R)$, is the definition of the probability density function (PDF).

$$d(z; A, R) = \frac{1}{2R} \exp\left(-\frac{|z-A|}{R}\right), -\infty \leq z \leq \infty. \quad (6)$$

The following is the Laplace distribution function:

$$\begin{aligned} d(z) &= \frac{1}{2} \exp\left(-\frac{|z-A|}{R}\right), z \leq A, \\ & 1 - \frac{1}{2} \exp\left(-\frac{|z-A|}{R}\right), z \geq A, \end{aligned} \quad (7)$$

where the location is shown by A , and the scale parameters are determined by R .

III.3 CLASSROOM BEHAVIOR RECOGNITION

Classroom behavior recognition is vital component of educational technology, aiming to understand and evaluate students' engagement, participation, and attentiveness during classroom activities. Accurate recognition of classroom behaviors provides valuable insights for educators to tailor instruction, provide timely feedback, and create supportive learning environments. The domain adaptive deep belief network (DA-DBN) is a sophisticated machine learning model designed to address the challenges of classroom behavior recognition.

DA-DBN combines principles from deep learning, unsupervised learning, and domain adaptation to learn robust representations of classroom behaviors across diverse environments and conditions. At its core, DA-DBN consists of multiple layers of interconnected neurons, organized into a hierarchical structure inspired by human brain's neural networks. DA-DBN incorporates domain adaptation mechanisms to align feature distributions between dissimilar provinces, mitigating the effects of sphere shift and improving the model's generalization presentation across diverse classroom environments. A DBN consists of multiple RBMs, each with a visible v layer and a hidden layer. RBM uses z vectors and units to connect layers within the same layer. The power function is defined by reflecting the state of the network, and the steady state of the network corresponds to the minimum state of the power function.

$$e(v, i) = -\sum_{h=1}^a m_h v_h - \sum_{g=1}^b n_g i_g - \sum_{j=1}^b \sum_{h=1}^a v_h i_h z_{hg} \quad (8)$$

Here m and n are the bias trajectory of observable and concealed layers, respectively. The joint circulation and conditional spreading of each stratum can be distinct as follows.

$$X(v, i) = \frac{E^{-e(v, i)}}{\sum_{v, i} E^{-e(v, i)}} \quad (9)$$

$$X(v | i) = \frac{X(v, i)}{X(i)} = \frac{E^{-e(v, i)}}{\sum_v E^{-e(v, i)}} \quad (10)$$

$$X(i | v) = \frac{X(v, i)}{X(v)} = \frac{E^{-e(v, i)}}{\sum_i E^{-e(v, i)}} \quad (11)$$

Since there is no correlation between components in each stratum, the restricted likelihoods of units can be designed using the following calculations.

$$X(i_g = 1 | v) = \frac{X(i_g=1, v)}{X(i_g=1, v) + X(i_g=0, v)} = \sigma\left(\sum_{h=1}^a m_h + z_{hg} v_h\right) \quad (12)$$

$$X(v_g = 1 | i) = \frac{X(v_g=1, i)}{X(v_g=1, i) + X(v_g=0, i)} = \sigma\left(\sum_{g=1}^b n_g + z_{hg} i_g\right) \quad (13)$$

$$\sigma(p) = \frac{1}{1 + E^{-p}} \quad (14)$$

The functional kernel I_k is definite as a Hilbert planetary with k . For two prospectdeliveries x and y for all $F \in I_k$, the DA-DBN value among x and y is definite as the recreated Hilbert interval distance among the fitted mean standards of x and y . It is defined as the quadratic formula of DA-DBN between x and y .

$$c_K^2(x, y) = \|\Delta e_{p^T \sim x}[\phi(p^T)] - e_{p^T \sim y}[\phi(p^T)]\|_{I_k}^2 \quad (15)$$

where ϕ is feature charting. The mouth kernel k can be labeled as follows.

$$K = \sum_{U=1}^a \beta_U K_U : \sum_{U=1}^a \beta_U = 1, \beta_U \geq 0, \forall U \quad (16)$$

where β_U is an extension of the weight constraint that ensures the production of regular multifunction kernels and a is the sum of the feature seeds. Thus, the optimization objective of the entire system contains of two portions: the organization error l_D in the source province and the insight error l_C among the foundation and mark domains. The optimization impartial is given by L is compute as follows.

$$l = l_D + l_C = \text{MIN} \frac{1}{b_t} \sum_{h=1}^{b_T} G(\theta(p_h^t), q_h^T) + \lambda \sum_{L=L_1}^L c_K^2(C_T^L, C_s^L) \quad (17)$$

The time difficulty of directly scheming the inner produce between the source field and the target domain $O(b^2)$ is too tall to be used effectively in a deep knowledge model.

$$c_K^2 = (x, y) = \frac{2}{b_i} \sum_{h=1}^{b_i/2} j_K(w_h) \quad (18)$$

$$w_h^\Delta = (p_{2h-1}^T, p_{2h}^T, p_{2h-1}^s, p_{2h}^s) \quad (19)$$

$$j_K(w_h)^\Delta = K(p_{2h-1}^T, p_{2h}^T) + K(p_{2h-1}^s, p_{2h}^s) + K(p_{2h-1}^s, p_{2h}^s) - K(p_{2h}^s, p_{2h-1}^s) \quad (20)$$

The limitations of the perfect are fine-tuned by stochastic gradient descent (SGD) technique. The incline of perfect limitation receipts the form.

$$\nabla \theta^L = \frac{\partial G(w_h)}{\partial \theta^L} + \lambda \frac{\partial j_K(w_h^L)}{\partial \theta^L} \quad (21)$$

Fine change of a deep neural network effortlessly achieved using a supervised learning procedure. An input pick, the output is obtained from the upper layer of the neural network.

$$r(p) = \text{Max}(0, p + j * X) : Q \sim B[0, \sigma(p)] \quad (22)$$

where j is the sound issue, which is second-hand to controller the equal of additional noise. The algorithm 1 describes the working process of classroom behavior recognition using the DA-DBN.

Algorithm 1 Classroom behavior recognition using DA-DBN

Input: The input to the algorithm is the energy function.

Output: After the learning process the upper layer of network is obtained.

1. *Begin*

2. *Initialized the population*

3. *The energy function can be described as*

$$e(v, i) = -\sum_{h=1}^a m_h v_h - \sum_{g=1}^b n_g i_g - \sum_{j=1}^b \sum_{h=1}^a v_h i_h z_{hg}$$

4. *The quadratic formulation of MK-MMD among x and y*

$$c_K^2(x, y)^\Delta = \|e_{p^T \sim x}[\phi(p^T)] - e_{p^T \sim y}[\phi(p^T)]\|_{I_k}^2$$

5. *The feature kernel k can be labelled as*

$$K = \sum_{U=1}^a \beta_U K_U : \sum_{U=1}^a \beta_U = 1, \beta_U \geq 0, \forall U$$

6. *For L=1*

7. *While do*

8. *else*

9. *The gradient of perfect limitation takes the form*

$$\nabla \theta^L = \frac{\partial G(w_h)}{\partial \theta^L} + \lambda \frac{\partial j_K(w_h^L)}{\partial \theta^L}$$

10. *End if*

11. *The output is obtained from the upper layer of the neural network*

$$r(p) = \text{Max}(0, p + j * X) : Q \sim B[0, \sigma(p)]$$

12. *else*

13. *Update the final value*

14. *End*

IV. RESULTS AND DISCUSSION

In this fragment, we current the results examination of proposed and existing study hall conduct acknowledgment models. To approves ourLHHO+DA-DBN model utilizing the publically available datasets like COCO and PASCAL VOC. The experimental environment was demonstrated on Ubuntu 18.04 functioning scheme, with an Intel(R) Xeon(R) Sliver 4110 CPU@2.10HZ CPU formation, a Ge Force GTX 1080ti graphics card selected for calculation, 11 GB of video memory, and DLaendaPyTorch version 1.7.0 and the Python environment used to construct the runtime environment. The consequences of proposed LHHO+DA-DBN model is contrasted and the current condition of craftsmanship imitations, for example, Quick R-CNN [32], SSD [33], YOLOv5 [34] and further developed YOLOv5 [35] [32] [33] [34] [36] [31].In guidance to evaluate the benefit of the arranged models, the assessment pointers utilize the exactness, accuracy, review, mean normal accuracy (Guide), F-measure, and the amount of outlines per extra to detect the pictures to appraise the performance.

IV.1 DATASET DESCRIPTION

To discourse the lack of public datasets for detecting schoolroom students' behavioral states, this study manually curated a dataset using real laboratory cassettes. The dataset comprises four performance categories: listening, looking down, lying down, and standing. Images were removed from noted class videos at frame breaks, ensuring diverse student poses and actions for robust model training. In total, 10,826 labels were annotated, with an 80:20 split for exercise and testing datasets, correspondingly. Table 1 presents the description of dataset. To prevent Overfitting caused by imbalanced training data, particularly due to a larger number of objects in some behaviors, data augmentation techniques were employed. Specifically, since standing samples were relatively scarce and challenging to collect, data augmentation was applied to expand the number of standing pose samples.

This involved intercepting images of students in standing posture and applying augmentation such as horizontal flipping, brightness adjustment, and Gaussian noise addition to enhance the dataset's diversity. For model training, images were resized to a standard input size of 640×640 pixels. A warm-up phase with linear interpolation learning rate adjustment was implemented initially, followed by a cosine annealing strategy for further learning rate adjustments. The training process spanned 100 epochs, with a batch size of 16 epochs per batch. These strategies collectively contributed to the actualexercise of a deep learning classical for accurate classroom behavior detection. Figure 2 shows the taster test imageries from the dataset.



Figure 2: Sample test images from dataset (a) listening, (b) looking down (c) lying down (d) standing.
Source: Authors, (2026).

Table 1: Dataset description.

Class	Number of samples		
	Training	Testing	Total
Listening	3105	776	3881
Looking down	2742	685	3427
Lying down	1843	460	2303
Standing	972	243	1215
Total samples	8662	2164	10826

Source: Authors, (2026).

IV.2 COMPARATIVE ANALYSIS

Table 2 presents a comprehensive comparison of various classroom students' behavioral detection models with the "listening" class. In terms of accuracy, the LHHO+DA-DBN model achieves the highest accuracy of 93.707%. This represents a significant improvement over Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 10.962%, 8.221%, 5.481%, and 2.74%, respectively. Regarding precision, the LHHO+DA-DBN model demonstrates superior precision at 92.184%. This marks a substantial enhancement compared to Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 17.291%, 12.969%, 8.646%, and 4.323%, respectively. Similarly, the LHHO+DA-DBN model exhibits remarkable recall at 92.245%, surpassing Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 17.28%, 12.96%, 8.64%, and 4.32%, respectively. Moreover, the proposed model achieves outstanding F-measure of 92.215%, shown significant improvements over Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 17.286%, 12.964%, 8.643%, and 4.321%, respectively. Furthermore, in terms of mAP, the LHHO+DA-DBN model achieves the highest score of 91.063%, outperforming Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 16.05%, 12.038%, 8.025%, and 4.013%, respectively. For a visual representation of these results, refer to Figure 2, which illustrates the comparison between the proposed and existing classroom students' behavioral detection models for the "listening" class.

Table 2: Comparative analysis of various classroom students' behavioral detection models with “listening” class.

Models	Number of epochs				
	20	40	60	80	100
	Accuracy (%)				
Fast R-CNN	82.093	82.984	83.623	83.751	84.723
SSD	84.661	85.552	86.191	86.319	87.291
YOLOv5	87.229	88.120	88.759	88.887	89.859
Improved YOLOv5	89.797	90.688	91.327	91.455	92.427
LHHO+DA-DBN	92.365	93.256	93.895	94.023	94.995
	Precision (%)				
Fast R-CNN	75.916	75.985	76.207	76.418	76.696
SSD	79.901	79.970	80.192	80.403	80.681
YOLOv5	83.886	83.955	84.177	84.388	84.666
Improved YOLOv5	87.871	87.940	88.162	88.373	88.651
LHHO+DA-DBN	91.856	91.925	92.147	92.358	92.636
	Recall (%)				
Fast R-CNN	75.634	75.918	76.426	76.518	77.029
SSD	79.619	79.903	80.411	80.503	81.014
YOLOv5	83.604	83.888	84.396	84.488	84.999
Improved YOLOv5	87.589	87.873	88.381	88.473	88.984
LHHO+DA-DBN	91.574	91.858	92.366	92.458	92.969
	F-measure (%)				
Fast R-CNN	75.775	75.951	76.316	76.468	76.862
SSD	79.760	79.936	80.301	80.453	80.847
YOLOv5	83.745	83.921	84.286	84.438	84.832
Improved YOLOv5	87.730	87.906	88.271	88.423	88.817
LHHO+DA-DBN	91.715	91.891	92.256	92.408	92.802
	mAP (%)				
Fast R-CNN	75.947	76.241	76.369	76.629	77.049
SSD	79.601	79.895	80.023	80.283	80.703
YOLOv5	83.255	83.549	83.677	83.937	84.357
Improved YOLOv5	86.909	87.203	87.331	87.591	88.011
LHHO+DA-DBN	90.563	90.857	90.985	91.245	91.665

Source: Authors, (2026).

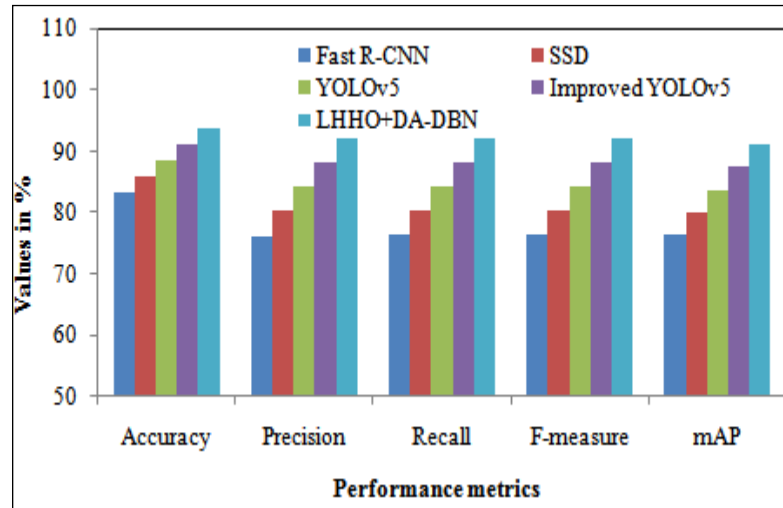


Figure 3: Results comparison for "listening" class.

Source: Authors, (2026).

The result analysis for Table 2, which compares various classroom students' behavioral detection models with the "Looking down" class: Our LHHO+DA-DBN model achieves the highest accuracy, reaching 97.365% after 100 epochs. Compared to the existing models, it outperforms Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 14.272%, 10.713%, 7.636%, and 4.468%, respectively. Similar to accuracy, the LHHO+DA-DBN model exhibits the highest precision, peaking at 95.478% after 100 epochs. It surpasses Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 12.988%, 9.741%, 6.494%, and 3.247%, respectively. Precision generally improves with the number of epochs for all models. LHHO+DA-DBN model also achieves the highest recall, reaching 94.898% after 100 epochs. Compared to Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5, it performs better by 16.416%, 13.370%, 8.225%, and 4.126%, respectively. Recall increases steadily with the number of epochs for all models. With an F-measure of 95.187% after 100 epochs, the LHHO+DA-DBN model again outperforms the other models. It surpasses Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 14.735%, 11.089%, 8.321%, and 4.625%, respectively. The LHHO+DA-DBN model achieves the highest mAP, reaching 93.235% after 100 epochs. It outperforms Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 14.028%, 11.439%, 7.178%, and 3.589%, respectively. For a visual representation of these results, refer to Figure 3, illustrates the comparison between the proposed and existing classroom students' behavioral detection models for the "looking down" class.

Table 3: Comparative analysis of various classroom students' behavioral detection models with "Looking down" class.

Models	Number of epochs				
	20	40	60	80	100
Accuracy (%)					
Fast R-CNN	81.360	81.740	82.573	83.185	83.093
SSD	84.928	85.308	86.141	86.753	86.661
YOLOv5	88.496	88.876	89.709	90.321	90.229
Improved YOLOv5	92.064	92.444	93.277	93.889	93.797
LHHO+DA-DBN	95.632	96.012	96.845	97.457	97.365
Precision (%)					
Fast R-CNN	81.535	81.247	81.868	82.024	82.490
SSD	84.782	84.494	85.115	85.271	85.737
YOLOv5	88.029	87.741	88.362	88.518	88.984
Improved YOLOv5	91.276	90.988	91.609	91.765	92.231
LHHO+DA-DBN	94.523	94.235	94.856	95.012	95.478
Recall (%)					
Fast R-CNN	77.482	77.622	77.855	78.172	78.395
SSD	81.608	81.748	81.981	82.298	82.521

YOLOv5	85.733	85.873	86.106	86.423	86.646
Improved YOLOv5	89.859	89.999	90.232	90.549	90.772
LHHO+DA-DBN	93.985	94.125	94.358	94.675	94.898
F-measure (%)					
Fast R-CNN	79.457	79.393	79.811	80.052	80.390
SSD	83.165	83.098	83.518	83.758	84.098
YOLOv5	86.866	86.797	87.220	87.458	87.800
Improved YOLOv5	90.562	90.491	90.915	91.153	91.496
LHHO+DA-DBN	94.253	94.180	94.606	94.843	95.187
mAP (%)					
Fast R-CNN	78.207	78.491	78.212	78.656	78.879
SSD	81.796	82.080	81.801	82.245	82.468
YOLOv5	85.385	85.669	85.390	85.834	86.057
Improved YOLOv5	88.974	89.258	88.979	89.423	89.646
LHHO+DA-DBN	92.563	92.847	92.568	93.012	93.235

Source: Authors, (2026).

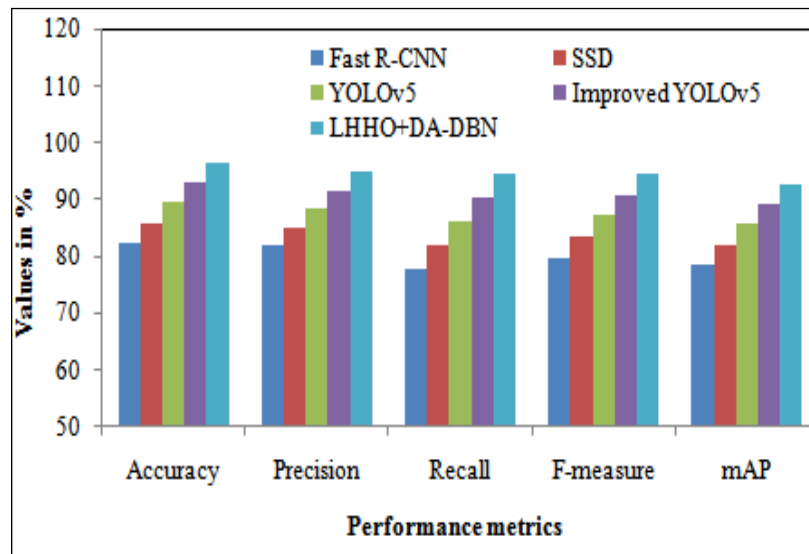


Figure 4: Results comparison for "Looking down" class.

Source: Authors, (2026).

Table 3 provides a comprehensive comparison of various classroom students' behavioral detection models with the "Lying down" class. The results showcase the performance of each model across different metrics and epochs. In terms of accuracy, the LHHO+DA-DBN model achieves the highest accuracy consistently across all epochs, reaching a maximum accuracy of 97.125%. It represents a significant improvement over other models, with the closest competitor being the Improved YOLOv5 model with maximum accuracy of 93.451%. Regarding precision, the LHHO+DA-DBN model also outperforms other models, achieving a precision of 95.235% at the highest. This is notably higher compared to other models, with the Improved YOLOv5 model coming closest with a precision of 91.248%. Similarly, in terms of recall, the LHHO+DA-DBN model shows superior performance, achieving a maximum recall of 95.023%.

It outperforms other models significantly, with the Improved YOLOv5 model having the closest recall at 89.385%. The F-measure, which considers both precision and recall, also highlights the effectiveness of LHHO+DA-DBN model. With maximum F-measure of 95.129%, it surpasses other models by a significant margin, indicating a balanced performance in terms of precision and recall. Lastly, the mAP metric, which evaluates the complete representation of the perfect, further confirms the superiority of the LHHO+DA-DBN model. With a maximum mAP of 94.022%, it demonstrates robustness and consistency across different epochs, outperforming other models in detecting the "Lying down" class behavior. For a visual representation of these results, refer to Figure 4, illustrates the comparison between the various classroom students' behavioral detection models for the "lying down" class.

Table 4: Comparative analysis of various classroom students' behavioral detection models with “Lying down” class.

Models	Number of epochs				
	20	40	60	80	100
Accuracy (%)					
Fast R-CNN	81.539	81.852	82.161	82.292	82.429
SSD	85.213	85.526	85.835	85.966	86.103
YOLOv5	88.887	89.200	89.509	89.640	89.777
Improved YOLOv5	92.561	92.874	93.183	93.314	93.451
LHHO+DA-DBN	96.235	96.548	96.857	96.988	97.125
Precision (%)					
Fast R-CNN	78.473	78.615	78.810	79.010	79.287
SSD	82.460	82.602	82.797	82.997	83.274
YOLOv5	86.447	86.589	86.784	86.984	87.261
Improved YOLOv5	90.434	90.576	90.771	90.971	91.248
LHHO+DA-DBN	94.421	94.563	94.758	94.958	95.235
Recall (%)					
Fast R-CNN	71.433	71.971	72.237	72.434	72.471
SSD	77.071	77.609	77.875	78.072	78.109
YOLOv5	82.709	83.247	83.513	83.710	83.747
Improved YOLOv5	88.347	88.885	89.151	89.348	89.385
LHHO+DA-DBN	93.985	94.523	94.789	94.986	95.023
F-measure (%)					
Fast R-CNN	68.770	69.111	69.341	69.540	69.697
SSD	75.128	75.469	75.699	75.898	76.055
YOLOv5	81.486	81.827	82.057	82.256	82.413
Improved YOLOv5	87.844	88.185	88.415	88.614	88.771
LHHO+DA-DBN	94.202	94.543	94.773	94.972	95.129
mAP (%)					
Fast R-CNN	74.946	75.038	75.105	75.239	75.402
SSD	79.601	79.693	79.760	79.894	80.057
YOLOv5	84.256	84.348	84.415	84.549	84.712
Improved YOLOv5	88.911	89.003	89.070	89.204	89.367
LHHO+DA-DBN	93.566	93.658	93.725	93.859	94.022

Source: Authors, (2026).

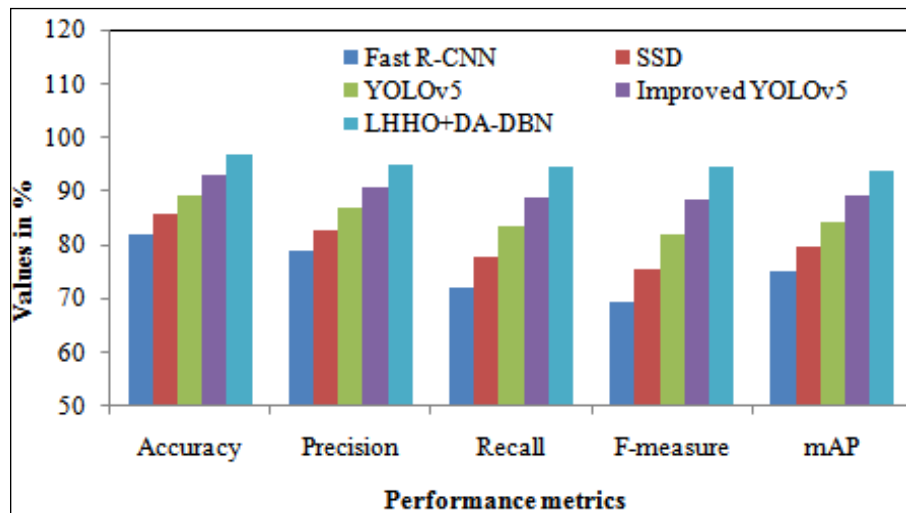


Figure 5: Results comparison for “Lying down” class.

Source: Authors, (2026).

Table 4 presents a comparative analysis of various classroom students' behavioral detection models with the "Standing" class. In terms of accuracy, the LHHO+DA-DBN model achieves the highest accuracy of 97.355% after 100 epochs, showing notable improvements over the other models such as Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 26.953%, 20.366%, 13.577%, and 7.589%, respectively.

Regarding precision, the LHHO+DA-DBN model achieves exceptional precision at 94.975%, outperforming the other models by margins of 15.948% (Fast R-CNN), 12.961% (SSD), 8.974% (YOLOv5), and 4.987% (Improved YOLOv5) after 100 epochs. Similarly, for recall, the LHHO+DA-DBN model demonstrates superior performance, achieving a recall of 96.525% after 100 epochs, surpassing Fast R-CNN, SSD, YOLOv5, and Improved YOLOv5 by 27.918%, 20.691%, 14.794%, and 7.877%, respectively. Additionally, in terms of F-measure and mAP, the LHHO+DA-DBN model consistently exhibits the highest values, indicating its effectiveness in accurately detecting students' "Standing" behavior in the classroom. Fig. 5 highlights the superior performance of the proposed LHHO+DA-DBN model for behavioral detection in the "Standing" class, emphasizing its potential for improving classroom monitoring and assessment.

Table 5: Comparative analysis of various classroom students' behavioral detection models with "Standing" class.

Models	Number of epochs				
	20	40	60	80	100
Accuracy (%)					
Fast R-CNN	70.402	70.54	70.633	70.666	70.999
SSD	76.991	77.129	77.222	77.255	77.588
YOLOv5	83.58	83.718	83.811	83.844	84.177
Improved YOLOv5	90.169	90.307	90.4	90.433	90.766
LHHO+DA-DBN	96.758	96.896	96.989	97.022	97.355
Precision (%)					
Fast R-CNN	78.077	78.31	78.71	78.947	79.027
SSD	82.064	82.297	82.697	82.934	83.014
YOLOv5	86.051	86.284	86.684	86.921	87.001
Improved YOLOv5	90.038	90.271	90.671	90.908	90.988
LHHO+DA-DBN	94.025	94.258	94.658	94.895	94.975
Recall (%)					
Fast R-CNN	68.017	68.281	68.399	68.557	68.937
SSD	74.914	75.178	75.296	75.454	75.834
YOLOv5	81.811	82.075	82.193	82.351	82.731
Improved YOLOv5	88.708	88.972	89.09	89.248	89.628
LHHO+DA-DBN	95.605	95.869	95.987	96.145	96.525
F-measure (%)					
Fast R-CNN	72.701	72.952	73.193	73.386	73.638
SSD	78.326	78.577	78.823	79.017	79.262
YOLOv5	83.877	84.127	84.379	84.574	84.812
Improved YOLOv5	89.368	89.617	89.874	90.070	90.303
LHHO+DA-DBN	94.808	95.057	95.318	95.516	95.744
mAP (%)					
Fast R-CNN	70.281	70.539	70.715	70.889	71.210
SSD	76.582	76.840	77.019	77.195	77.510
YOLOv5	82.831	83.088	83.272	83.448	83.759
Improved YOLOv5	89.037	89.293	89.480	89.657	89.964
LHHO+DA-DBN	93.154	93.285	93.645	93.859	93.988

Source: Authors, (2026).

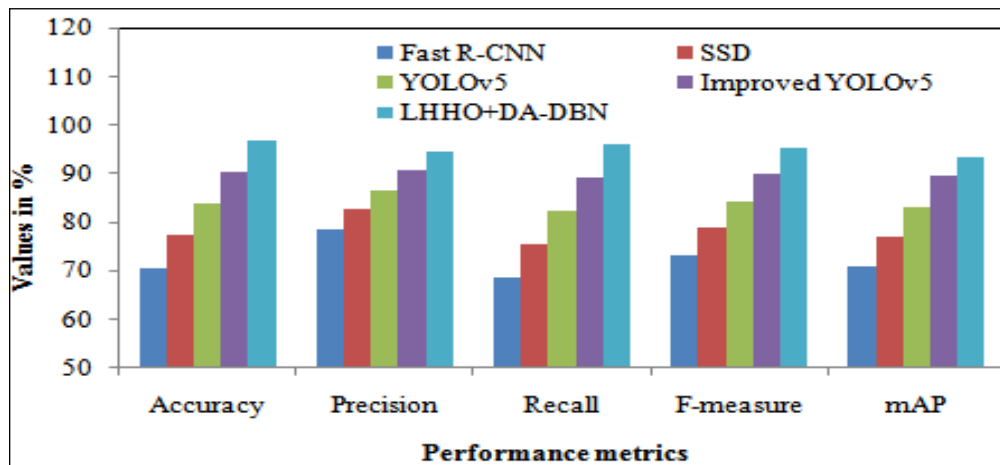


Figure 6: Results comparison for "Standing" class.

Source: Authors, (2026).

V. CONCLUSION

In this education, we introduced a novel method for enhancing classroom behavior recognition and evaluation among higher education students. Our method integrates a hybrid optimization and deep learning technique, leveraging a hybrid spatio-temporal ResNet combined with the Laplacian Harris hawk optimization (LHHO) algorithm for feature extraction and optimization. This innovative approach enables the extraction of crucial contextual, semantic, and positional features while effectively addressing challenges such as occluded object recognition. Moreover, we used a domain adaptive deep belief network (DA-DBN) specifically tailored for classroom behavior recognition tasks. By leveraging this network, we observed important developments in the detection correctness of occluded objects, crucial for accurately assessing student behaviors. To assess the efficacy of our technique, we conducted comprehensive tests using widely-used public datasets, including COCO and PASCAL VOC. The results of our evaluations validate the larger presentation of our proposed algorithm associated to current classroom behavior recognition models. Notably, the LHHO+DA-DBN model achieved an impressive average accuracy of 96.031% and a recall rate of 94.335%, consistently outperforming state-of-the-art models across various scenarios.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Janaki Sivakumar, Mehmet Aydin, Sudhanshu Maurya, M. PrasannaBlessy, Kirti Hemant Wanjale and S. Agalya.

Methodology: M. PrasannaBlessy, Kirti Hemant Wanjale and S. Agalya.

Investigation: Janaki Sivakumar, Mehmet Aydin, Sudhanshu Maurya.

Discussion of results: Janaki Sivakumar, Mehmet Aydin, Sudhanshu Maurya, M. PrasannaBlessy, Kirti Hemant Wanjale, S. Agalya.

Writing – Original Draft: M. PrasannaBlessy, Kirti Hemant Wanjale and S. Agalya.

Writing – Review and Editing: Kirti Hemant Wanjale and S. Agalya.

Resources: Janaki Sivakumar and Sudhanshu Maurya.

Supervision: Janaki Sivakumar and Sudhanshu Maurya.

Approval of the final text: Janaki Sivakumar, Mehmet Aydin, Sudhanshu Maurya, M. PrasannaBlessy, Kirti Hemant Wanjale, S. Agalya.

VII. REFERENCES

- [1] Wang, W., 2024. Application of deep learning algorithm in detecting and analyzing classroom behavior of art teaching. *Systems and Soft Computing*, p.200082.
- [2] Zhao, G., Zhang, Y. and Chu, J., 2024. A multimodal teacher speech emotion recognition method in the smart classroom. *Internet of Things*, 25, p.101069.
- [3] Liu, G.L. and Wang, Y., 2024. Modeling EFL teachers' intention to integrate informal digital learning of English (IDLE) into the classroom using the theory of planned behavior. *System*, 120, p.103193.
- [4] Dang, M., Liu, G., Xu, Q., Li, K., Wang, D. and He, L., 2024. Multi-object behavior recognition based on object detection for dense crowds. *Expert Systems with Applications*, p.123397.
- [5] Hong, C., 2024. Group Behavior Recognition Algorithm for Basketball Video Based on Hidden Tag Mining Strategy under Gamification Teaching and Learning. *Entertainment Computing*, p.100638.
- [6] Foster, J.K., Korban, M., Youngs, P., Watson, G.S. and Acton, S.T., 2024. Automatic classification of activities in classroom videos. *Computers and Education: Artificial Intelligence*, p.100207.
- [7] Spielberg, J., Burkhardt, T., Carreon, E.D. and Gitlow, E.R., 2024. Fostering healthy social and emotional climates in early childhood classrooms through infant and early childhood mental health consultation. *Early Childhood Research Quarterly*, 67, pp.307-319.
- [8] Zhao, J. and Zhu, H., 2023. Cbph-net: A small object detector for behavior recognition in classroom scenarios. *IEEE Transactions on Instrumentation and Measurement*.
- [9] Sümer, Ö., Goldberg, P., D'Mello, S., Gerjets, P., Trautwein, U. and Kasneci, E., 2021. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*.
- [10] Li, L., Liu, M., Sun, L., Li, Y. and Li, N., 2022. ET-YOLOv5s: toward deep identification of students' in-class behaviors. *IEEE Access*, 10, pp.44200-44211.
- [11] Bhattacharjee, A. and Das, S., 2020. D2SC-GAN: dual deep-shallow channeled generative adversarial network, for resolving low-resolution faces for recognition in classroom scenarios. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(3), pp.223-234.
- [12] Mahalakshmi, V., Sandhu, M., Shabaz, M., Keshta, I., Prasad, K.D.V., Kuzieva, N., Byeon, H. and Soni, M., 2024. Few-shot learning-based human behavior recognition model. *Computers in Human Behavior*, 151, p.108038.
- [13] Akila, D., Garg, H., Pal, S. and Jeyalakshmi, S., 2023. Research on recognition of students attention in offline classroom-based on deep learning. *Education and Information Technologies*, pp.1-29.
- [14] Mo, J., Zhu, R., Yuan, H., Shou, Z. and Chen, L., 2023. Student behavior recognition based on multitask learning. *Multimedia Tools and Applications*, 82(12), pp.19091-19108.
- [15] Huang, Y., Peng, C. and Liu, S., 2023. Empirical Research of Classroom Behavior Based on Online Education: A Systematic Review. *Mobile Networks and Applications*, pp.1-13.

- [16] Mandia, S., Singh, K. and Mitharwal, R., 2023. Recognition of student engagement in classroom from affective states. *International Journal of Multimedia Information Retrieval*, 12(2), p.18.
- [17] Chen, Z., Liang, M., Xue, Z. and Yu, W., 2023. STRAN: Student expression recognition based on spatio-temporal residual attention network in classroom teaching videos. *Applied Intelligence*, 53(21), pp.25310-25329.
- [18] Gu, M., Feng, J. and Chu, Y., 2023. A novel multi-scale facial expression recognition algorithm based on improved Res2Net for classroom scenes. *Multimedia Tools and Applications*, pp.1-18.
- [19] Zhou, J. and Herencsar, N., 2023. Abnormal Behavior Determination Model of Multimedia Classroom Students Based on Multi-task Deep Learning. *Mobile Networks and Applications*, pp.1-14.
- [20] Pabba, C., Bhardwaj, V. and Kumar, P., 2023. A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom. *Multimedia Tools and Applications*, pp.1-31.
- [21] Chen, S. and Chen, Z., 2020. On human behavior recognition with deep learning and IR spectral signal restoration technologies in a natural classroom. *Infrared Physics & Technology*, 105, p.103167.
- [22] Liu, T., Wang, J., Yang, B. and Wang, X., 2021. Facial expression recognition method with multi-label distribution learning for non-verbal behavior understanding in the classroom. *Infrared Physics & Technology*, 112, p.103594.
- [23] Sun, B., Wu, Y., Zhao, K., He, J., Yu, L., Yan, H. and 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, pp.8335-8354.
- [24] Burleigh, C.L. and Wilson, A.M., 2021. Teachers' awareness in identifying microaggressive behaviors within the K-12 classroom. *Social Psychology of Education*, 24(1), pp.143-167.
- [25] Lin, F.C., Ngo, H.H., Dow, C.R., Lam, K.H. and Le, H.L., 2021. Student behavior recognition system for the classroom environment based on skeleton pose estimation and person detection. *Sensors*, 21(16), p.5314.
- [26] Chen, H. and Guan, J., 2022. Teacher-student behavior recognition in classroom teaching based on improved YOLO-v4 and Internet of Things technology. *Electronics*, 11(23), p.3998.
- [27] Thamilarasi V., Roselin. R., "Application of Machine Learning in chest X-Ray images", International book & River publisher series in computing and information science and Technology, pp: 52-1,2023. ISBN 9788770228114, 9788770228107.
- [28] Yin Albert, C.C., Sun, Y., Li, G., Peng, J., Ran, F., Wang, Z. and Zhou, J., 2022. Identifying and Monitoring Students' Classroom Learning Behavior Based on Multisource Information. *Mobile Information Systems*, 2022.
- [29] Chen, H., Zhou, G. and Jiang, H., 2023. Student Behavior Detection in the Classroom Based on Improved YOLOv8. *Sensors*, 23(20), p.8385.
- [30] Shou, Z., Yan, M., Wen, H., Liu, J., Mo, J. and Zhang, H., 2023. Research on Students' Action Behavior Recognition Method Based on Classroom Time-Series Images. *Applied Sciences*, 13(18), p.10426.
- [31] J. Anand, V. Thamilarasi, A. Rayal, H. K. Gupta, K. Jyothi and P. Vishwakarma, "Fuzzy Logic-Based Deep Learning for Human-Machine Interaction and Gesture Recognition in Uncertain and Noisy Environments," 2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT), Bhimtal, Nainital, India, 2025, pp. 435-440, doi: 10.1109/CE2CT64011.2025.10939160.
- [32] Tang, L., Xie, T., Yang, Y. and Wang, H., 2022. Classroom Behavior Detection Based on Improved YOLOv5 Algorithm Combining Multi-Scale Feature Fusion and Attention Mechanism. *Applied Sciences*, 12(13), p.6790.
- [33] Girshick, R. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
- [34] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y. and Berg, A.C., 2016. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14* (pp. 21-37). Springer International Publishing.
- [35] Wang, Z., Li, L., Zeng, C. and Yao, J., 2023. Student learning behavior recognition incorporating data augmentation with learning feature representation in smart classrooms. *Sensors*, 23(19), p.8190.
- [36] Bochkovskiy, A., Wang, C.Y. and Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.