IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

An automated recognition of teacher and student activities in the classroom environment: A deep learning framework

RAJAMANICKAM YUVARAJ[1], A. AMALIN PRINCE [2], M. MURUGAPPAN[3,4]

[1]Science of Learning in Education Centre (SoLEC), Office of Education Research, National Institute of Education, Nanyang Technological University, Nanyang Walk, 637616, Singapore

[2]Department of Electrical and Electronics Engineering, BITS Pilani, Goa Campus, Sancoale, 403726, Goa, India.

[3]Intelligent Signal Processing (ISP) Research Lab, Department of Electronics and Communication Engineering, Kuwait College of Science and Technology, Block 4, Doha 13133, Kuwait.

[4]Department of Electronics and Communication Engineering, Faculty of Engineering, Vels Institute of Sciences, Technology, and Advanced Studies, Chennai 600117, Tamilnadu, India.

*Corresponding author. A Amalin Prince, Department of Electrical and Electronics Engineering, BITS Pilani, Goa Campus, Sancoale, 403726, Goa, India. Email: amalinprince@goa.bits-pilani.ac.in

ABSTRACT Teacher and student behavior during class is often observed by education professionals to evaluate and develop a teacher's skill, adapt lesson plans, or monitor and regulate student learning and other activities. Traditional methods rely on accurate manual techniques involving in-person field observations, questionnaires, or the subjective annotation of video recordings. These techniques are time-consuming and typically demand observation and coding by a trained professional. Thus, developing automated tools for detecting classroom behaviors using artificial intelligence could greatly reduce the resources needed to monitor teacher and student behaviors for research, practice, or professional development purposes. This paper presents an automated framework using a deep learning approach to recognize classroom activities encompassing both student and teacher behaviors from classroom videos. The proposed method utilizes a long-term recurrent convolutional network (LRCN), which captures the spatiotemporal features from the video frames. For evaluation purposes, experiments were carried out on a subset of the EduNet and an independent dataset composed of classroom videos collected from the internet. The proposed LRCN system achieved a maximum average accuracy (ACC) of 93.17%, precision (PRE) of 94.21%, recall (REC) of 91.76%, and F1-Score (F1-S) of 92.60% on EduNet dataset when estimated by 5-fold cross-validation. The system provides ACC = 83.33%, PRE = 89.25%, REC = 83.32%, and F1-S = 82.14% when applied to independent testing which ensures reliability. The study has significant methodological implications for the automated recognition of classroom activities and may assist in providing information about classroom behaviors that are worthy of inclusion in the evaluation of education quality.

INDEX TERMS education, classroom activities, teacher, student, machine learning, LRCN, deep learning.

I. INTRODUCTION

Artificial intelligence (AI) is currently used in a wide range of aspects of our lives, making it a popular field of research among researchers worldwide. AI has made significant advances, and researchers are currently using AI to identify human actions in many fields, including cooking [1], sports [2], [3], and everyday activities [4]. The use of mobile devices, surveillance cameras, and CCTV cameras has increased in recent years, resulting in large amounts of data in the form of videos uploaded occasionally to online media platforms like YouTube. Consequently, AI research has focused on detecting and recognizing activities within these online videos and camera feeds, leveraging the capabilities of big data and digital technologies to advance activity recognition and analysis.

In recent years, educational psychology and pedagogy researchers have used AI methods to evaluate the quality of education. Information about the student-teacher interaction is crucial for measuring the quality of education in a classroom. In addition to reminding students to regulate their behavior, recognizing classroom action helps teachers improve their teaching methods. Their behavioral information reflects the

classroom atmosphere, which is valuable for determining students' learning styles, psychological characteristics, and personalities and creating better lesson plans. Classroom behaviors are traditionally recorded through field observations and questionnaire surveys, but they require much time and effort. Classrooms nowadays are equipped with CCTV cameras, which capture massive amounts of video, allowing AI technology to automatically recognize student and teacher activities and monitor active class participation. Little progress has been made in the field of automatic student and teacher action recognition from video, particularly in the classroom. Automated methods of detecting classroom behaviors using machine learning algorithms could reduce the resources required for monitoring teacher and student behavior for research, practice, or professional development.

To date, most machine learning (ML)-based tools for video-recorded behaviour recognition have been developed for analysing actions outside of education contexts [1], [4]. ML tools have rarely been developed or tested for recognizing classroom behavior by teachers or students. A summary of recent studies on automatic classroom behavior recognition from videos is presented in Table 1. Using an end-to-end system, Sun et al. (2021) found that student's behavior in the classroom could be detected, identified, and captioned [5]. A ResNet-101 network was used to extract spatial and temporal information about behavior from video frames. Students' activity recognition system based on deep convolutional generative adversarial networks (DCGAN) was developed by [6]. Student classroom behaviors such as standing, sitting, climbing on the table, writing, using a smartphone, raising a hand, and gazing out of the window can all be detected [6]. It achieves an accuracy rate of 98% across all categories. Using Zernike moment of motion images in combination with optical flow, Wu et al. (2016) developed a technique to identify hand-ups, standing-ups, and sitting-downs during class [7].

A feed-forward learning model combined with an extreme learning machine (ELM) classifier has been proposed by Nida et al. (2019) to recognize different instructor behaviors in the classroom [8]. In their article [9], Ren et al. (2002) developed an algorithm to recognize human actions within smart classrooms. In their paper, the authors describe a system that allows teachers to recognize complex actions. It locates two shoulders in a silhouette image and uses them to locate basic motion features, such as elbows, hands, and faces, using a 2-order B-spline function. Then, detection of natural context-dependent actions is then performed using a primitive-based coupled hidden Markov model. Gang et al. (2021) developed a system to analyze large amounts of teaching videos using teaching data sets based on teacher behavior [10]. In the proposed method, there was an educational pattern called teacher set, which is described as the area in the classroom where the teacher should be present. Afterward, a more advanced behavior recognition algorithm is applied, which uses 3D bilinear pooling and is capable of representing 3D features more effectively, making it possible to identify teachers based on their behavior. A sequential recognition method based on hidden Markov models (HMM) was proposed by Race et al. (2015) that recognizes five natural teacher actions in a lecture room: writing, walking, pointing to the board, pointing to a presentation, and standing, with a recognition rate of more than 90% [11].

In an effort to develop better classroom models that can handle the true complexity of classroom environments, limited studies have attempted to capture teacher and student behavior together. Chen et al. (2022) propose a system for teacher-student behavior recognition based on YOLO version 4 and Internet of Things (IoT) [12]. The proposed algorithm uses the IOT paradigm to obtain classroom data in the perceptron layer and then sends that data to the processing layer for target detection using DL and Single Shot. A greater than 90% accuracy rate was recorded for identifying sustainable classroom behaviors. Sharma et al. (2021) implemented a model combining Inflated 3D (I3D) and ResNet-50 (i.e., I3D-ResNet-50), which could automatically recognize 20 unique actions of both teacher and student in live classroom environments and achieved 72.3% overall accuracy [[13]].

In [21], the authors used ten CCTV video recordings of teacher activities for an hour at 25 frames per second. They have classified 11 different actions of the teachers in the class, including cleaning, pointing, standing, talking, writing, presentation cleaning, presentation pointing, presentation standing, presentation talking, presentation writing, and writing and talking, and achieved a maximum mean classification rate of 94% using 3D-CNN. Muhammad et.al also used CCTV recordings of four activities, namely, class with the teacher, no class (grouping of students), exam with the instructor, and no exam (grouping of students in the exam hall) using CNN [20]. In their study, video frames of various sizes were analyzed (224×224, 128×128, 64×64, and 32×32), and the smaller frame size (32×32) resulted in 99% accuracy. However, the authors have not tested their algorithms against other open-source datasets to benchmark their performance. In a recent experiment, Wang and his team used three motion sensors to conduct fourteen classroom activities on 13 subjects [22]. A combination of recurrent neural networks (RNN) and convolutional neural networks (CNN) has been used for classification, including LSTM and BiLSTM, as well as 1D-CNN and DNN. By using 1D-CNN, they achieved a maximum mean accuracy of 100%, while by using BiLSTM, they achieved 99.8% accuracy.

Afsana et.al used voice signals to classify classroom activities into three types, single voice signal, multiple voice signals, and no voice, in 2022 [23]. Their study used both recurrent neural networks (LSTM) and deep neural networks (DNN) as well as convolutional neural networks, and achieved 100% accuracy using the LSTM. A recent study by Foster et al utilized their own dataset collected from classroom activities of elementary school children over a period of 1000 hours to classify the activities based on an improved Background Suppression Network (Bas-Net+) [24]. A total of four different activities have been considered by the

TABLE 1. Summary of classroom action recognition studies using ML algorithms from videos. The studies are presented in reverse chronological order.

Author	Target population	Data used	No. of classroom actions	Classroom Actions	Methodology	Results summary
Chen et al. (2022)	Students and teacher (classroom)	500 min videos	9	It includes asking questions, both hand gestures, referring to the projection, no hand gestures, guidelines for raising the hand, walking around, writing on the blackboard, and other similar behaviors.	Improved YOLO-v4	ACC = 90% (average)
Sun et al. (2021)	Students (classroom)	128 videos	11	There are a number of behaviors that can be observed such as listening attentively, taking notes, using a mobile phone, yawning, eating, drinking, looking around, using a computer, and snoring.	Reconstruction network (Recent)	ACC = 73.50%
Gang et al. (2021)	Teacher (classroom)	340 videos	8	Interacting with students can be done in a variety of ways, including asking them questions, pointing to the blackboard, writing on it, cleaning it, using an interactive whiteboard, inviting them to answer, or using a calculator.	3D bilinear pooling network	ACC = 81.00%
Sharma et al. (2021)	Students and teacher (classroom)	7851 videos	20	A classroom argument, clap, eat, introduce the topic, gossip, raise your hand, hit, hold the book, hold the mobile phone, hold the stick, read the book, sit on the chair, sit at the desk, slap, sleep, stand, talk, walk in the classroom, write on the board, and write on your textbook.	I3D-ResNet-50	ACC = 72.30%
Cheng et al. (2020)	Students (simulated)	400	7	The students are walking, sitting, climbing on the table, writing, playing with their smartphones, holding or raising their hands, and looking around.	DCGAN	ACC = 98.00%
Nida et al. (2019)	Instructor (lecture room)	100 videos from IAVID	9	Interacting or being idle, pointing to the board or screen, reading notes, using a mobile phone, using a laptop, sitting, walking, or writing on the board.	CNN+ELM	ACC = 81.25%
Wu et al. (2016)	Students (classroom)	91	3	Hand-up actions, stand up, and sit down	Lucas-Kanade optical flow	ACC = 84.54%
Raza et al. (2015)	Instructor (lecture room)	50 videos	5	Writing, walking, point to board, point to presentation, and standing	GMM + HMM	ACC = 90.00%
Ren et al. (2002)	Teacher (smart classroom)	50 videos	7	The act of taking objects from the desk and returning them to it, pointing at the blackboard, communicating with the students, explaining objects, and drinking water are all part of the teaching process.	Motion features + PCHMM	ACC = 90.00%

DCGAN - Deep Convolutional Generative Adversarial Network; HMM - hidden Markov model; PCHMM - primitive-based coupled hidden Markov model; GMM - Gaussian mixture model; IAVID - Instructor Activity Video Dataset; ACC – Accuracy.

researchers, such as whole group activities, individual activities, small group activities, and transitions. Their proposed methodology resulted in a recognition rate of 88% for whole-class activities, 84% for small-group activities, 89% for individual activities, and 93% for transition activities. Fan Yang and his team developed one of the largest spatio-temporal image datasets with 757265 images from kindergarten to high school classroom videos [25]. There are three behaviors in the dataset: hand-raising, reading, and writing. Additionally, they developed a multi-model fusion based on slow-fast algorithms, YOLOv5, YOLOv7, and YOLOv8 to classify classroom behaviors with an average accuracy of 82.3%. Jia et.al have recently combined YOLOv5 with contextual attention (CA) mechanisms to accurately recognize student behavior in the classroom [26]. They collected the samples from surveillance videos with a maximum duration of 550 minutes, generating a dataset with five different activities (raising the hand, standing up, writing, and slipping). Furthermore, the CA mechanism was used in conjunction with VGG-16 networks to improve behavior detection accuracy. As a result, they achieved an average precision of 82.1% using the VGG-16 model.

These state-of-the-art investigations illustrate the potential for ML techniques to automatically capture and recognize teacher, student, and teacher+student behaviors from video recordings of realistic classroom activities. However, while these few studies provide important advances in ML classroom behavior recognition, they provide limited validation and reporting of the computational complexity of the applied techniques. The development of a robust and reliable classroom behavior monitoring system remains a valuable yet challenging problem.

A. THE PRESENT STUDY

The body of research dedicated to developing and applying machine learning (ML) techniques for human behavior recognition is indeed extensive, with significant contributions from studies such as [18], [1], and [19]. These studies have advanced our understanding of behavior recognition across various contexts, employing sophisticated ML models to identify and interpret human actions in diverse scenarios. In previous works, motion sensors, voice signals, spatiotemporal images, and video recordings have been used to distinguish different classroom activities. The earlier works, however, used their own datasets, did not compare their results with open-source data, used a limited number of subjects (samples), and deep neural networks were the most preferred types of models. There remains a notable gap in research specifically focused on applying ML techniques to behavior recognition within classroom learning environments for a smaller dataset. This gap is particularly critical given the unique challenges posed by educational settings,

where contextual factors such as varying teaching styles, classroom dynamics, and diverse student interactions play a significant role in shaping behavior. ML models, while powerful, often require context-specific training to achieve optimal performance.

In classroom settings, the models need to be trained on data that accurately represents the complexity and variety of classroom interactions. Generic models developed for other environments may not translate effectively to the classroom without appropriate adaptation and testing. The current shortage of research in this area underscores the need for studies that specifically focus on training and testing ML models within realistic classroom environments. To address these gaps the present study draws upon the EduNet dataset [13] to investigate an automated ML framework for recognizing several teacher and student behaviors recorded during classroom lessons (e.g., talking, reading, writing on board, hand-raising). The core ML model examined in this study is the long-term recurrent convolutional network (LRCN), which is mainly used to learn features from both spatial and temporal information to process sequential data. The LRCN effectively integrates spatial details with motion-related information, which is crucial for analyzing long sequences such as videos. Its capability to handle both types of data enables it to capture and interpret complex patterns across extended video frames, making it an appropriate choice for video analysis tasks.

II. RESEARCH MATERIALS AND METHODS

Here we present a detailed picture of the classroom video dataset used, the deep learning methodology for recognizing classroom activities, including pre-processing, as well as the description of the proposed LRCN model's structure and key parameters. We also provide a brief overview of the evaluation process. Fig 1 details the methodological framework of the proposed automated pipeline for classroom action recognition.

A. CLASSROOM DATASET

This study analysed a subset of the EduNet dataset obtained from [13] upon the request of the authors. The characteristics of EduNet and the datasets used in this study are summarized in Table 2. The EduNet dataset contains 7851 video clips featuring twenty different teacher and student actions, respectively. The majority of video clips were recorded in real classroom environments in twelve schools, while others were sourced from YouTube. All video clips had a frame rate (FR) of 30 frames per second (FPS), a resolution of 1280 × 780, and were manually annotated with a class action label (e.g., hand-raise). For this study, a subset of EduNet was utilized; this subset, referred to henceforth as the "study dataset", contained 927 video clips with durations ranging from 1–14 seconds. The study dataset features ten different action classes, including four teacher actions (holding a mobile phone, explaining the subject, writing on a board, and holding a book) and six student actions (arguing, eating, hand-raising, reading a book, sitting at a desk, and writing in a textbook). Examples of teacher and student behaviours captured in the study dataset are shown in Figure 2.

The performance of the proposed automated framework was evaluated using an independent dataset comprised of classroom videos collected from the iStock website (https://www.istockphoto.com/) [14]. Hereafter, this dataset is referred as "independent dataset". Five iStock videos were collected for each teacher and student action. The videos ranged from 4–30 seconds long, with a mean FR of 28.37 FPS and a uniform resolution of 640 × 360 pixels. The search terms used to identify relevant iStock videos were the action class names in EduNet (e.g., "arguing") and their synonyms (e.g., "disagreeing"). All fifty videos identified for evaluation were watched by the research team to confirm their relevance to their respective action class. A comparison of the number of video clips for each action in the EduNet dataset with those in the study dataset is shown in Figure 3.

B. PRE-PROCESSING

Several pre-processing steps were used to prepare annotated classroom video data. This included frame extraction, frame selection, and resizing. In each video, there are a large number of frames that are almost identical to one another. Thus, the first twenty video frames are processed per second with a time interval of 0.66 seconds in this frame selection step. This reduces computational overhead and maintains a uniform input to the LRCN, as the videos have varying durations. OpenCV libraries were used to extract the video frames [15]. After the frame selection, all action images were resized to 64 × 64 (width × height) pixels and then passed to LRCN for training, validation, and evaluation.

C. LRCN DEEP NEURAL NETWORK METHODOLOGY

An LRCN is a machine learning model that combines convolutional and recurrent neural networks [16], [17]. The model analyzes sequential data, such as videos and audio, by capturing both geographical and temporal patterns. LRCN has been used to develop several video processing applications, including voice recognition, captioning, and video action recognition. The goal of this study was to build a spatiotemporal deep learning model that recognizes the actions of students and teachers using CNNs and LSTMs. Figure 4 shows the process of developing the classroom action recognition model. The action image input is denoted as F_i $(i = 1, 2, 3, 4, \ldots n)$, where n represents the frame numbers. A CNN extracts action features from image sets, producing a fixed-length feature vector fv_i (i=1,2,3,4,. . . .l). Deep-learned features are then broken up into sequential components and passed on to repetitive LSTMs. Finally, the output is sent to a fully connected layer to recognize classroom actions. Our deep learning algorithms and analyses were developed and conducted using the Keras framework and TensorFlow backend on a computer equipped with a GeForce 1070 graphics card, 7th generation Intel core, 32GB

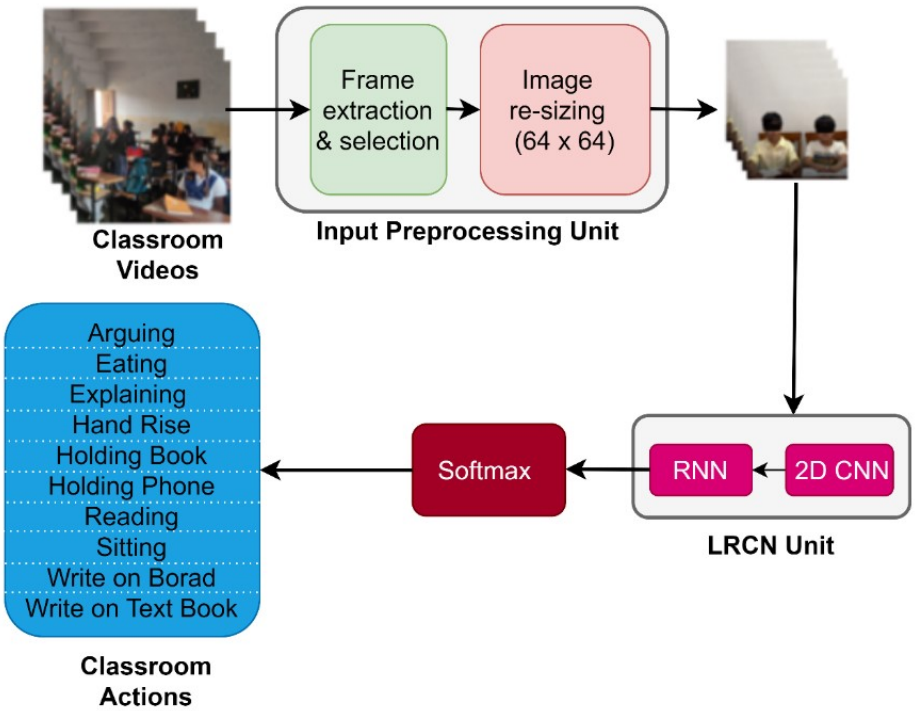

FIGURE 1. Block diagram for our suggested algorithmic pipeline for classroom action recognition.

TABLE 2. Detailed information on EduNet, the study dataset, and the independent dataset.

Specifications of video data	EduNet	Study dataset	Independent dataset
Number of videos in total	7851	927	50
Number of student action video clips	4228	443	30
Number of teacher action video clips	3623	484	20
Number of classroom actions in total	20	10	10
Number of student actions	9	6	6
Number of teacher actions	11	4	4
Video duration (range in seconds)	3.25 to 12.7	1 to 14	4-30
Total duration	12 hours	1.628 hours	11.53 minutes

of RAM, and a CPU of 3.20 GHz running Windows 10 Home (64-bit).

1) CNN-based feature extraction

In this study, a 2D-CNN was used with input dimensions of 64×64 $(width \times height)$. As shown in Table 3, there are five convolution layers in the CNN. As a first step, we implemented a 5×5 convolution kernel and a 2×2 maximum (max) pooling to preserve significant features. Three other convolution layers (layers 3,5,7, and 9) have kernel sizes of 3×3, strides of 1×1, and maximum pooling over a 2×2 region. The accuracy of the test improved as there were more convolution kernels used. The convolution kernels in these layers are 16, 32, 64, 128, and 256. Each convolution layer was normalized using BatchNorm normalization method, with a leaky ReLU with a coefficient of 0.01. By reducing neuron distribution variation and increasing the learning rate, BatchNorm normalizes each small batch to a zero mean and unit variance [17].

2) Sequential learning by LSTM

The output from the last max-pool was fed to the LSTM layer. We performed experiments on the LSTM architectures shown in Figure 5, starting with a single-layer LSTM network and moving to more complex ones. It was found that LSTM networks with a single layer had a less significant effect than networks with a double layer. The performance of the LSTN network improved less when there were more

FIGURE 2. Examples of teacher and student action in a classroom setting used in this study (top and bottom).

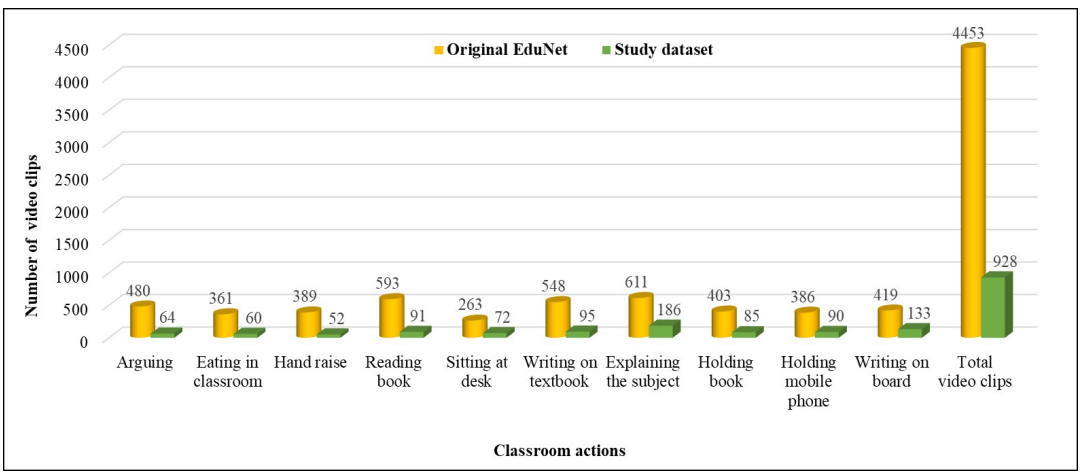

FIGURE 3. Amount of video clips for each action class in the EduNet and Study datasets.

layers added. Therefore, we used two layers in the LSTM network, and 512 units were initially configured. Prior to the dense layer, a flattened layer was applied as the dense layer dimensions needed to be one-dimensional. The first dense layer is composed of 512 neurons, which are then connected to the dropout layer with a rate parameter of 0.25 to reduce overfitting. Then, a second dense layer with "n" neurons performs recognition into the different classroom action classes. The neuron number varies for three different groups: teacher-centric actions (4 classes), student-centric actions (6 classes), and student + teacher-centric actions (i.e., classroom) (10 classes). Finally, SoftMax activation is applied to the output. Table 3 presents the complete details of the proposed LRCN model layer parameters.

Based on the validation procedure (see section 2.4), hyperparameters like epoch numbers and batch size are chosen. Several combinations of hyperparameters were tested based on a heuristic approach to determine the most effective combination to improve the performance. As a consequence of this process, the hyperparameters selected for this study included a learning rate of 0.0001, an Adam optimizer with momentum equal to 0.9, as well as categorical cross-entropy as a loss function to train the LSTM model. The model was trained for 50 epochs in a batch size of four to achieve the highest classification accuracy.

3) Performance evaluation

The scheme for training, validation, and testing in this study can be seen in Figure 6. Based on recognition accuracy (ACC), precision (PRE), recall (REC), and F1-score (F1-S), we evaluated the proposed LRCN-based deep learning framework. In addition, confusion matrices from the test fold data were used to visualize the model's correct recognitions and incorrect recognitions. The 5-fold cross-validation method was adopted in our study to ensure a consistent recognition performance. During this process, the study dataset was divided into five equal subsets (almost equal in some folds) randomly, ensuring consistency in distribution. Each subset is repeated five times, with one forming the test set and the other four forming the training set. Each fold training set used 90% of the data for training the LRCN model and 10% for tuning hyperparameters. A test fold (final model) is then made based on the fold that produces the

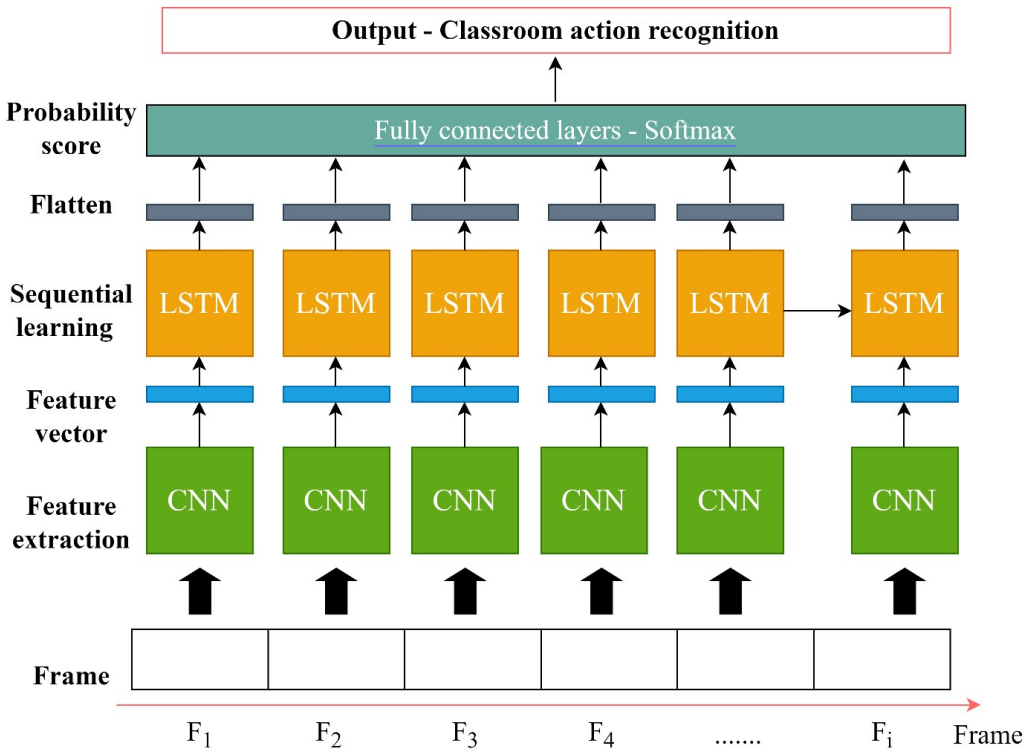

FIGURE 4. An overall framework for our proposed classroom action recognition system using LRCNs. Three other convolution layers (layers 3, 5, 7, and 9) have kernel sizes of 3×3, strides of 1×1, and maximum pooling over a 2×2 region. The accuracy of the test improved as more convolution kernels were used. The convolution kernels in these layers are 16, 32, 64, 128, and 256. Each convolution layer was normalized using BatchNorm normalization method, with a leaky ReLU with a coefficient of 0.01. By reducing neuron distribution variation and increasing the learning rate, BatchNorm normalizes each small batch to a zero mean and unit variance [17].

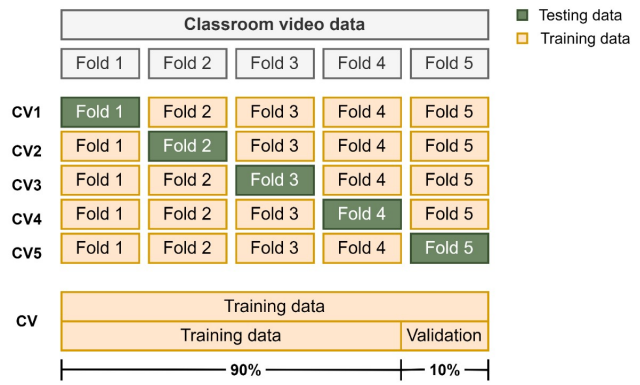

FIGURE 5. Training, validation, and testing scheme. CV denotes the cross-validation.

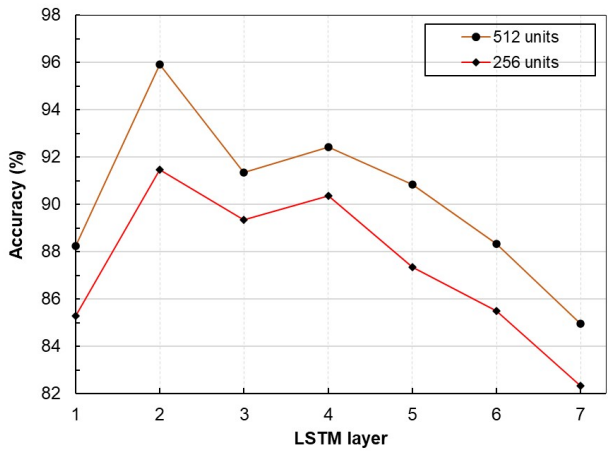

FIGURE 6. Comparison of LSTM networks using different layers to recognize classroom behavior.

most performance. In order to evaluate the overall classroom activity recognition performance, the average of the ACC, PRE, REC, and F1-S was calculated across all folds. A final analysis of average performance was conducted for three groups: teacher actions (4 action classes: explaining the subject, holding the book, holding a mobile phone, and writing on the board), student actions (6 action classes: arguing, eating in the classroom, raising hands, reading books, sitting at desks, and writing on textbooks), and classroom actions (10 classes).

III. RESULTS AND DISCUSSION

In this section, the results of the proposed LRCN model are presented on the study dataset as well as on an independent dataset. In Figure 7, ACC is shown during the training phase.

TABLE 3. Complete details of proposed LRCN neural network architecture.

Layer No.	Hidden layer	Filter No.	Kernel size	Stride	Others
0	Input (64 x 64)	-	-	-	-
1	2DConv1 + LeakyReLu + BatchNorm1	16	5 x 5	1	-
2	Max Pooling1	-	2 x 2	2	-
3	2DConv2 + LeakyReLu + BatchNorm2	32	3 x 3	1	-
4	Max Pooling2	-	2 x 2	2	-
5	2DConv3 + LeakyReLu + BatchNorm3	64	3 x 3	1	-
6	Max Pooling3	-	2 x 2	2	-
7	2DConv4 + LeakyReLu + BatchNorm4	128	3 x 3	1	-
8	Max Pooling4	-	2 x 2	2	-
9	2DConv5 + LeakyReLu + BatchNorm5	256	3 x 3	1	-
10	Max Pooling5	-	2 x 2	2	-
11	LSTM + BatchNorm	512	-	-	-
12	LSTM + BatchNorm	512	-	-	-
13	Flatten	-	-	-	-
14	Dense1	512	-	-	-
15	Dropout	-	-	-	Rate = 0.25
16	Dense2	-	-	-	n_class

TABLE 4. Classroom action recognition performances obtained using the proposed LRCN model of the study dataset. Best performed fold is highlighted in bold in each group.

Classroom actions	Folds	ACC (%)	PRE (%)	REC (%)	F1-S (%)
Teacher	Fold 1	85.21	86.52	84.82	85.63
	Fold 2	91.68	92.34	90.57	91.77
	Fold 3	92.32	93.01	92.17	**92.57**
	Fold 4	89.53	90.51	92.23	89.84
	Fold 5	**95.37**	**96.01**	**94.04**	95.52
	Average	90.37 ± 3.77	91.68 ± 3.50	90.77 ± 3.54	91.07 ± 3.66
Student	Fold 1	88.25	87.51	89.51	88.25
	Fold 2	92.95	90.32	91.74	92.98
	Fold 3	91.21	90.44	91.01	91.22
	Fold 4	89.42	88.34	90.88	89.44
	Fold 5	**94.42**	**94.01**	**94.99**	**94.50**
	Average	91.26 ± 2.51	90.12 ± 2.51	91.63 ± 2.05	91.28 ± 2.54
Classroom	Fold 1	91.29	93.03	92.23	91.34
	Fold 2	93.87	**95.68**	92.13	93.35
	Fold 3	93.94	94.26	91.96	**94.06**
	Fold 4	90.82	91.94	89.53	90.23
	Fold 5	**95.91**	96.14	**92.95**	94.00
	Average	93.17 ± 2.10	94.21 ± 1.76	91.76 ± 1.30	92.60 ± 1.72

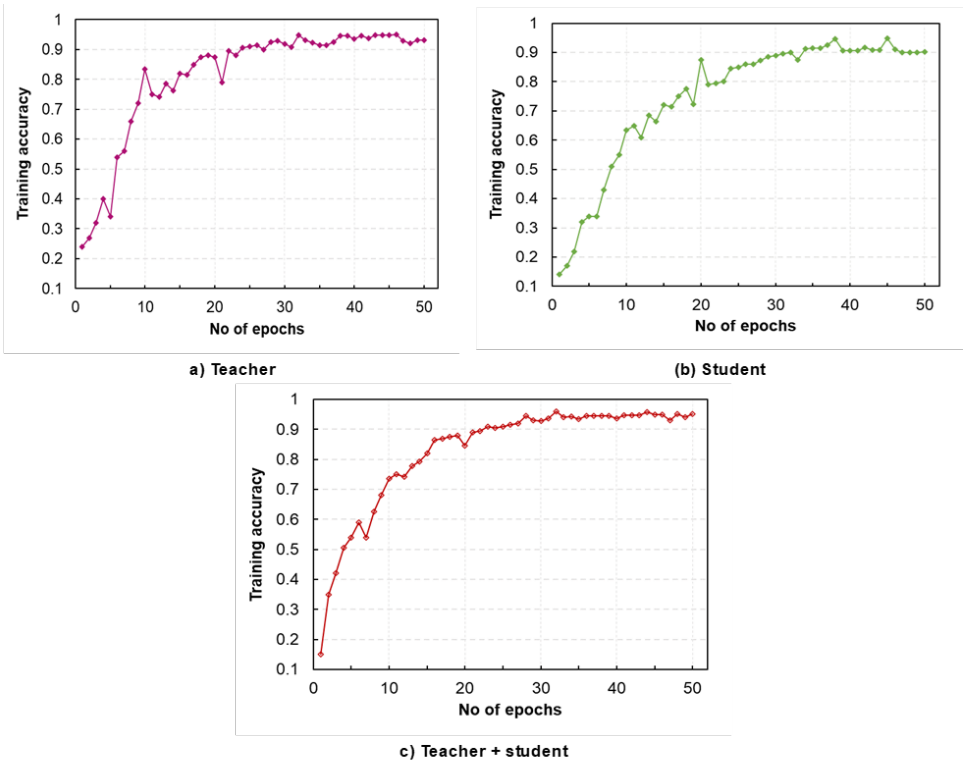

FIGURE 7. Accuracy during LRCN training process. (a) Teacher actions, (b) Student actions, and (c) student + teacher actions.

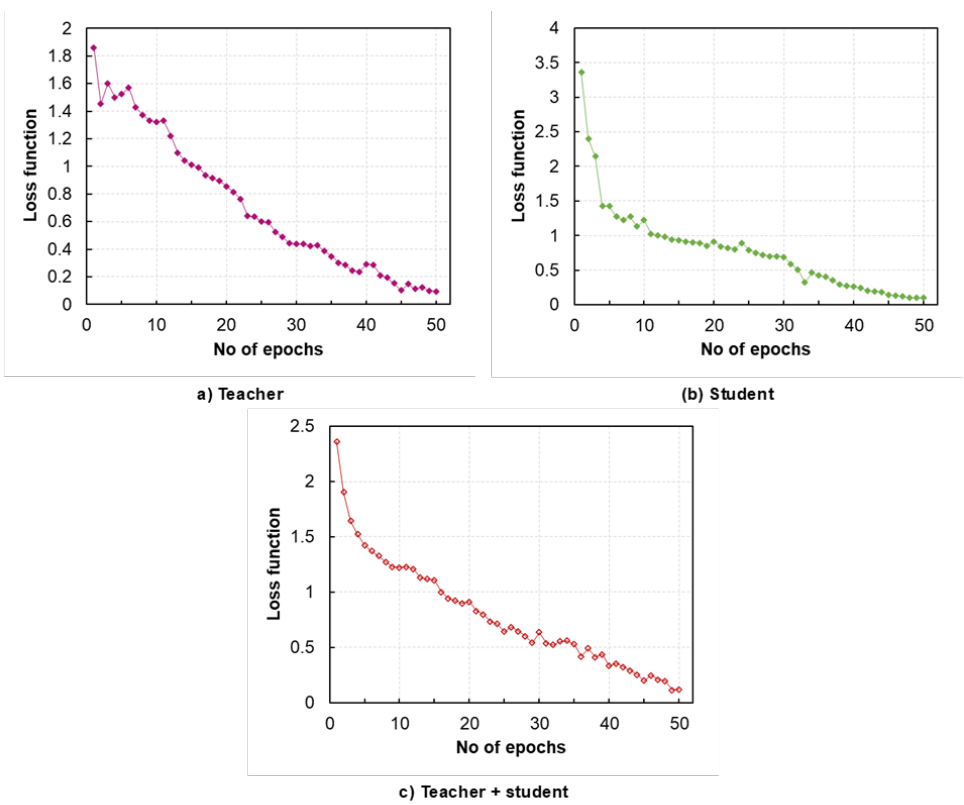

FIGURE 8. Loss function during LRCN training process. (a) Teacher actions, (b) Student actions, and (c) student + teacher actions.

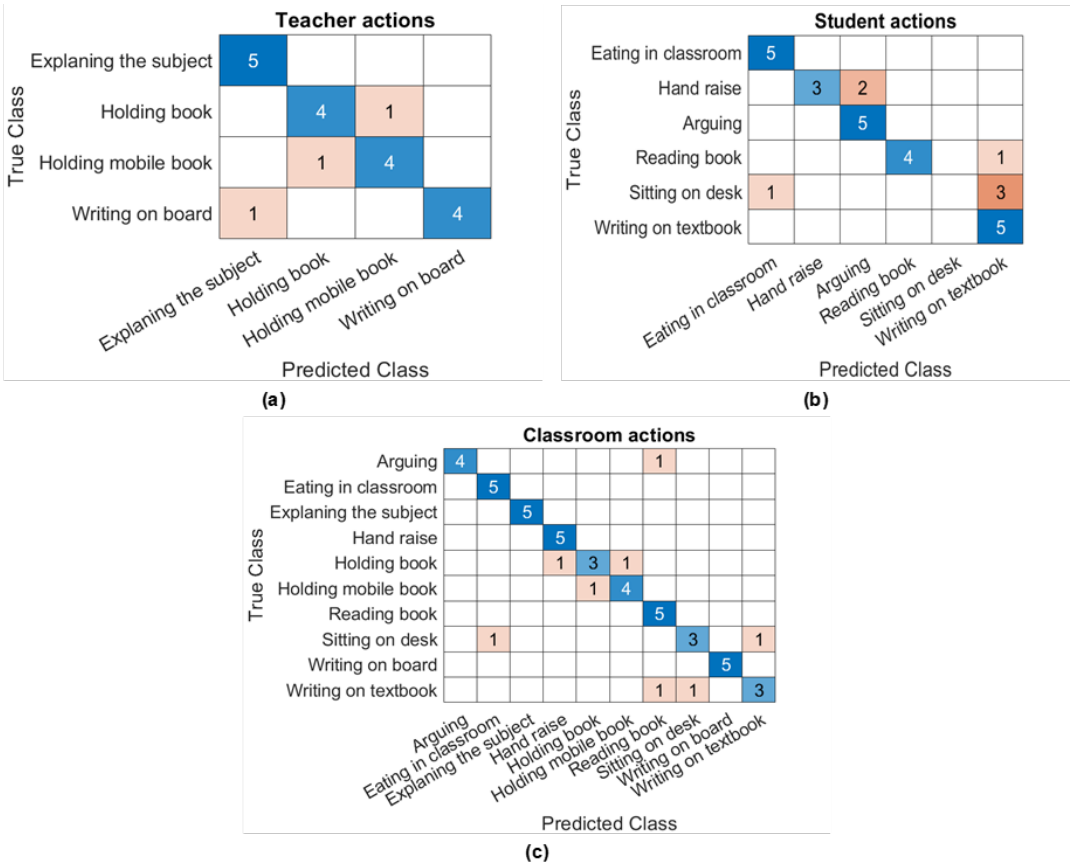

FIGURE 9. Confusion matrix on an independent dataset (a) Teacher actions, (b) Student actions, and (c) Classroom actions.

TABLE 5. Computational time cost evaluation of each step averaged over 25 classroom videos.

Task	Time cost (s)
Loading video file from hard drive	0.092 ± 0.028
Loading LRCN model from hard drive	0.298 ± 0.201
Video data pre-processing	1.139 ± 0.927
LRCN evaluation (CPU + GPU)	0.163 ± 0.114
Total computational time	**1.692**

TABLE 6. Performance evaluation results of LRCN-based classroom action recognition model on an independent dataset. All metrics are given as %.

Group	No of actions	No. of video clips for testing	ACC	PRE	REC	F1-S
Classroom	10	50	84.00	88.33	84.00	84.38
Teacher	4	20	85.00	90.74	85.00	84.82
Student	6	30	83.33	89.25	83.32	82.14

In Figure 7, we see that the ACC for the model increases as new epochs are added during training. A rapid increase in ACC was observed during the early epochs. In general, a value of approximately 0.8 results in a slower increase in accuracy, but it continues to increase. A teacher's, students, and teacher + student's action is estimated to be approximately 93.00%, 90.00%, and 95.00% accurate in training. In Figure 8, we see a decrease in loss function as training progresses. There is a direct correlation between the loss function and ACC, indicating good recognition results. The first epoch of training is marked by a rapid decline in values. As the loss function decreases around 0.3, it reaches its lowest values of approximately 0.09, 0.09, and 0.02 for teacher, student, and teacher-student actions, respectively. A decrease in the loss function implies an increase in accuracy due to a reduction in errors. It is generally true that accuracy increases as loss decreases (but not always).

Table 4 displays the performance of the proposed LRCN model, showing that the model had relatively high performance in terms of ACC ($\geq$ 93%), PRE ($\geq$ 94), REC ($\geq$ 92), and F1-S ($\geq$ 94%). For the teacher-centric actions, the LRCN model yielded an average ACC = 90.82%, PRE = 91.68%, REC = 90.77%, and F1-S = 91.07%. For student-centric actions, the model delivered an average ACC = 91.26%, PRE = 90.12%, REC = 91.63%, and F1-S = 91.28%. For the teacher + student actions, an average ACC = 93.17%, PRE = 94.21%, REC = 91.76%, and F1-S = 92.60% were achieved. Higher ACC, PRE, REC, and F1-S values are obtained for all the groups. The satisfactory performance of the LRCN model indicates that a LRCN model can perform well for small databases. The proposed LRCN model also delivered performance with the lowest standard deviation (SD) of accuracy, showing greater consistency. Overall, the

LRCN achieved higher recognition, and generalization was also observed in the three groups. Based on the results of this study (see Table 1), the proposed model is significantly more effective than other methods currently available in the literature at capturing spatial and temporal information from input video frames, in addition to the dynamics within those frames, compared to other methods.

Table 5 presents the time profiling of the LRCN model for recognizing a classroom action. The computational time was obtained after averaging the time computed for 25 videos of 7-second duration from the best performed fold (See Table 4) of the classroom videos of this study dataset. These 25 videos were chosen because the maximum number of videos were present for the 7-second duration in the fold. In total, the proposed LRCN system takes approximately two seconds (on a CPU + GPU system) to recognize the classroom actions. This computational speed may be sufficient for online feedback to understand the classroom behaviour during lessons, which may be used as feedback to assist students in regulating their behaviour, help teachers improve or adjust their method of instruction and potentially evaluate teaching performance.

In order to demonstrate the reliability and potential generalizability of the proposed LRCN approach, an independent dataset was used for classroom activity recognition. Using the collected videos, LRCN-based classroom action recognition testing was performed. The results of the classroom action recognition performances on an independent dataset are summarized in Table 6. For student-centric actions, it can be noticed from the table that the LRCN model provides ACC = 83.33%, PRE = 89.25%, REC = 83.32%, and F1-S = 82.14% applied to an independent dataset. For the teacher-centric actions, the model achieved ACC = 85.00%, PRE = 90.74%, REC =85.00%, and F1-S = 84.82%. For the classroom actions, the proposed LRCN model yielded an ACC = 84.00%, PRE = 88.33%, REC = 84.00%, and F1-S = 84.38%. These results clearly demonstrate the generalizability of the proposed approach and might fit well with the new classroom dataset. Figure 9 displays the confusion matrix of our result on an independent dataset, where the developed model recognized almost all the actions (about 83%) correctly, and only a small number of samples were recognized incorrectly.

A. LIMITATIONS AND FUTURE DIRECTIONS

This study has a few major limitations. First, we utilized only 20% of the EduNet dataset for our classroom activity recognition experiments, which may limit the robustness and generalizability of our proposed scheme. Consequently, our approach has not been fully validated with the complete dataset. In future work, we aim to leverage the entire EduNet dataset to comprehensively validate our methodology. Additionally, although our system is designed to recognize classroom activities effectively, it has not yet been tested in real-time, online settings within actual classrooms. Moving forward, we plan to implement and evaluate online recognition of classroom activities, which will provide valuable insights into the system's real-world applicability. We also intend to expand our research by incorporating a broader range of classroom actions and analyzing a more extensive collection of real classroom videos.

Based on the recent works on classroom activities recognition, Muhammad et al. has used different sizes (224×224, 128×128, 64×64, and 32×32) of the images to classify four types of classroom activities and achieved a maximum mean classification rate at the resolution of 32×32 [20]. However, we recognize that using 64×64 dimensions for input images is unconventional in traditional recognition tasks, where higher resolutions are typically employed. Hence, the use of 64×64 images for input images is another limitation of our study, which is unconventional compared to traditional recognition tasks that typically use higher resolutions. As a result of this smaller size, significant details may be lost, potentially impacting the model's performance and ability to capture finer details. In light of our hardware limitations, we chose this resolution to balance computational efficiency and processing time. In future work, we plan to evaluate larger image sizes, such as 224×224 or 128×128, to overcome this limitation. As a result, we will be able to understand the trade-offs between image resolution, computational efficiency, and recognition performance.

However, to address this limitation, we plan to evaluate larger image sizes in future work, such as 224×224 or 128×128. This will help us understand the trade-offs between image resolution, computational efficiency, and recognition performance. While our framework demonstrates effective performance using current deep learning techniques, it has not been compared directly with Transformer-based models, which have shown significant advancements in handling sequential and complex data. We will incorporate such comparisons in future research by utilizing the full EduNet dataset. Specifically, we intend to explore how Transformer-based approaches could be integrated or benchmarked against our existing framework to evaluate potential improvements in recognition accuracy and overall performance. Finally, the proposed model's generalizability across diverse classroom settings and varying video qualities requires further consideration. The classroom environment, including size, layout, and teaching style, may affect the model's predictions, particularly in contexts not well-represented in the training data.

Additionally, video quality fluctuations—such as lighting conditions, camera angles, and resolution—may introduce variability in model performance. Preprocessing methods like noise reduction or standardization may mitigate some of these effects, but the model's sensitivity to lower-quality inputs poses limitations. Further research should explore adaptation strategies, such as fine-tuning, to enhance the model's robustness and ensure consistent performance across a wider range of educational contexts. Also, we could analyze small facial expression variations within a DNN by using a deep efficient face network [27]. We could also integrate a hybrid learning mode identification framework into machine learning for identifying different types of educational envi-

ronments [28].

While significant advances in AI have enabled human action detection across diverse domains like cooking, sports, and daily activities, the application of deep learning (DL) architectures in classroom environments remains relatively underexplored. This work presents an initial step in adapting well-established DL techniques to the unique context of classroom behavior analysis. Although broader integration of classroom actions and comprehensive automation are not fully addressed in this study, the aim was to demonstrate the feasibility and potential of this approach. Future work will extend this analysis by incorporating a deeper comparison with other algorithms on the same dataset and developing fully automated systems. The outcomes of this paper, therefore, contribute to the literature as a foundational approach, upon which more advanced and holistic methods can be built.

IV. CONCLUSIONS

This study aims to develop a method that uses machine learning to automatically classify the classroom behaviour of teachers and students from video recordings taken in real classroom environments. A deep learning framework using LRCN was proposed to recognize classroom actions in an automated way. We evaluated the LRCN model using EduNet data with annotated classroom videos that featured typical student and teacher behavior. The model was also tested on an independent classroom video to ensure its reliability and generalizability. For teacher + student-centric actions, the LRCN model achieved an average maximum accuracy (ACC) of 93.17%, precision (PRE) of 94.21%, recall (REC) of 91.76%, and F1-Score of 92.60%. With this high level of performance, this automated framework could have significant methodological implications for the automated recognition of classroom activities and provide valuable information regarding classroom behaviors that can be used to evaluate education quality. Furthermore, the system might enable teachers to understand student behavior in classrooms to reveal their learning styles, and track teacher actions to gain a more holistic understanding of the classroom environment. The outcomes of this study also provide a research basis for using AI to solve educational problems, such as analyzing and developing automated tools, which is beneficial for related research, e.g., pedagogy and educational psychology. The major contributions of this work are: (i) the development of an automated pipeline for classroom activity recognition using the LRCN deep learning framework, (ii) the achievement of state-of-the-art performance using the EduNet dataset, (iii) the evaluation of the model's performance on an independent dataset of classroom video footage.

ACKNOWLEDGMENT

The authors would like to thank the research teams who collected and made the datasets available publicly and for granting access to the datasets.

REFERENCES

[1] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, OH, USA, 2014, pp. 780-787.

[2] Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L. Large-Scale video classification with convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE; 2014. p. 1725–32

[3] Rodriguez MD, Ahmed J, Shah M. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE; 2008. p. 1-8.

[4] Heilbron FC, Escorcia V, Ghanem B, Niebles JC. ActivityNet: A large-scale video benchmark for human activity understanding. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA2015.

[5] Sun B, Wu Y, Zhao K, He J, Yu L, Yan H, et al. Student Class Behavior Dataset: a video dataset for recognizing, detecting, and captioning students' behaviors in classroom scenes. Neural Computing and Applications. 2021;33(14):8335-54.

[6] Cheng Y, Dai Z, Ji Y, Li S, Jia Z, Hirota K, et al. Student action recognition based on deep convolutional generative adversarial network. Chinese Control And Decision Conference (CCDC). Hefei, China: IEEE; 2020.

[7] Wu D, Dang D, Wang J. Recognition of students combining features of Zernike moment and optical flow. IEEE International Conference on Computer and Communications (ICCC): IEEE; 2016. p. 676-9.

[8] Nida N, Yousaf MH, Irtaza A, Velastin SA. Instructor activity recognition through deep spatiotemporal features and feedforward extreme learning machines. Mathematical Problems in Engineering. 2019:13.

[9] Ren H, Xu G. Human action recognition in smart classroom. Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition. Washington, DC, USA: IEEE; 2002.

[10] Gang Z, Wenjuan Z, Biling H, Jie C, Hui H, Qing X. A simple teacher behavior recognition method for massive teaching videos based on teacher set. Applied Intelligence. 2021;51:8828-49.

[11] Raza A, Yousaf MH, Sial HA, Raja G. HMM-based scheme for smart instructor activity recognition in a lecture room environment. Smart Computing Review. 2015;6(6):578-90.

[12] Chen H, Guan J. Teacher–Student behavior recognition in classroom teaching based on improved YOLO-v4 and internet of things technology. Electronics. 2022;11(3998):1-15.

[13] Sharma V, Gupta M, Kumar A, Mishra D. EduNet: A New Video Dataset for Understanding Human Activity in the Classroom Environment. Sensors. 2021;21(17):5699.

[14] iStock. https://wwwistockphotocom/.

[15] OpenCV. https://opencvorg/.

[16] Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. arXiv:14114389v4. 2016.

[17] Wei X, Zhou L, Zhang Z, Chen Z, Zhou Y. Early prediction of epileptic seizures using a long-term recurrent convolutional network. J Neurosci Methods. 2019;327:108395.

[18] Heilbron, FC., Escorcia, V., Ghanem, B., Niebles, JC. (2015). ActivityNet: A large-scale video benchmark for human activity understanding. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA

[19] Rodriguez, M. D., Ahmed, J., Shah, M. (2008). Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). Anchorage, AK, USA: IEEE.

[20] M. Wasim, I. Ahmed, J. Ahmad and M. M. Hassan, "A Novel Deep Learning Based Automated Academic Activities Recognition in Cyber-Physical Systems," in IEEE Access, vol. 9, pp. 63718-63728, 2021, doi: 10.1109/ACCESS.2021.3073890.

[21] Rafique MA, Khaskheli F, Hassan MT, Naseer S, Jeon M (2022) Employing automatic content recognition for teaching methodology analysis in classroom videos. PLoS ONE 17(2): e0263448. https://doi.org/10.1371/journal. pone.0263448

[22] Wang, H.; Gao, C.; Fu, H.; Ma, C.Z.-H.; Wang, Q.; He, Z.; Li, M. Automated Student Classroom Behaviors' Perception and Identification Using Motion Sensors. Bioengineering 2023, 10, 127. https://doi.org/10.3390/bioengineering10020127

[23] Mou, A., Milanova, M., Baillie, M. Active Learning Monitoring in Classroom Using Deep Learning Frameworks. In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13643. Springer, Cham.

[24] Foster, J.K., Korban, M., Youngs, P., Watson, G.S., Acton, S.T. (2024). Automatic classification of activities in classroom videos. Comput. Educ. Artif. Intell., 6, 100207.

[25] Yang, Fan and Xiaofei Wang. "Student Classroom Behavior Detection based on Spatio-Temporal Network and Multi-Model Fusion.", arXiv, 2023, https://doi.org/10.48550/arXiv.2310.16267

[26] Jia, Q.; He, J. Student Behavior Recognition in Classroom Based on Deep Learning. Appl. Sci. 2024, 14, 7981. https://doi.org/10.3390/app14177981

[27] Hoang, T. M., Nam, G. P., Cho, J., Kim, I. J. (2020). "Deface: Deep Efficient Face Network for Small Scale Variations." IEEE Access, 8, 142423-142433.

[28] Verma, C., Illés, Z., Kumar, D. (2024). "TCLPI: Machine Learning-Driven Framework for Hybrid Learning Mode Identification." IEEE Access, 12, 98029-98045.

RAJAMANICKAM YUVARAJ is a Research Scientist at the National Institute of Education (NIE), Nanyang Technological University (NTU), Singapore. He received both B.E. (Electronics and Communication Engineering) and M.E. degrees (Biomedical Engineering-Gold Medalist) from Anna University and a Ph.D. degree (Biomedical Electronics) from University Malaysia Perlis (UniMAP), Malaysia, 2015. Before joining NIE, he was a Postdoctoral Research Fellow in the School of Electrical and Electronic Engineering at NTU (2017–2020) and the Neural Systems Lab, Department of Biomedical Engineering at the University of Kentucky, USA. His primary research interests include machine learning algorithms, with applications to biomedical signal analysis that include affective computing in education, educational neuroscience, pattern recognition, mental health, and well-being.

A. AMALIN PRINCE (Senior Member, IEEE) is a Professor and Head of the Department of Electrical and Electronics Engineering, Birla Institute of Technology and Science, Pilani, Goa Campus, India. He received his M.E. in Applied Electronics from the Sathyabama Institute of Science and Technology, Chennai, India, in 2005 and his Ph.D. from the Birla Institute of Technology and Science, Pilani, Rajasthan, India, in 2011. He His research interests include FPGA-based System Design, Hardware Accelerated Data Processing, Affective Computing, and Applications of Artificial Intelligence.

M. MURUGAPPAN has been working at Kuwait College of Science and Technology (KCST), Kuwait as a Full Professor in Electronics, Department of Electronics and Communication Engineering since 2016. He is also serving as a Visiting Professor at the School of Engineering at Vels Institute of Science, Technology, and Advanced Studies in India and an International Visiting Fellow at the Center of Excellence in Unmanned Aerial Systems at Universiti Malaysia Perlis in Malaysia. In 2006, he graduated from Anna University, India with an M.E. degree in Applied Electronics. He received his Ph.D. from Universiti Malaysia Perlis, Malaysia in 2010 for his contribution to the field of Mechatronic Engineering. Between 2010 and 2016, he worked as a Senior Lecturer at the School of Mechatronics Engineering, Universiti Malaysia Perlis, Malaysia. In a study by Stanford University, he was recently ranked among the top 2 percent of scientists working in experimental psychology and artificial intelligence for three consecutive years (2020-2022). To date, his google scholar citations reaches 7000+ with an H-Index of 42. His research into affective computing has received more than 750K in grants from Malaysia, Kuwait, and the UK. His publications include more than 140 peer-reviewed conference proceedings papers, journal articles, and book chapters. Several of his journal articles have been recognized as best papers, best papers of the fiscal year, etc. Prof. Murugappan is a member of the editorial boards of PlosOne, Human Centric Information Sciences, PEERJ Computer Science, Journal of Medical Imaging and Health Informatics, and International Journal of Cognitive Informatics. In addition to being the Chair of the IEEE Kuwait Section's Educational Activities Committee, Kuwait. He is interested in affective computing, affective neuroscience, neuromarketing, and medical image.