Videos Classification based on the Behaviors of Children in Pre-school through Surveillance Cameras

Nguyen Tran Gia The[1,2], Do Pham Phuc Tinh[1,2], Cao Dinh Duy Ngoc[1,2], Nguyen Huu Minh Tam[1,2], Ngo Huynh Truong[1,2], and Do Trong Hop[1,3]

[1] University of Information Technology, VNU-HCM, Vietnam
[2] {20521940,20522020,20521661,20521871,20522085}@gm.uit.edu.vn
[3] hopdt@uit.edu.vn

Abstract. Human actions classification is one of the research fields in computer vision. Recent advancements in image processing using machine learning and deep learning methods have significantly improved the ability to classify objects from images and videos. Efficient action classification models can be used in many applications such as video surveillance, human-machine interaction, and sports analysis. There are few datasets that exist in the school domain, especially in preschool.Therefore, in this study, we introduce the **B**ehaviors of **C**hildrens **i**n **P**re-**S**chool (**BCiPS**), a new dataset for the task of recognizing and classifying abnormal activities in the preschool domain. We collected videos from surveillance cameras in the classroom, helping to recognize abnormal actions of children in time and resolve incidents. The dataset consists of 4.268 videos with a length of 3-6 seconds and a total size of 19.9GB. We use methods to extract frames based on motion and then apply neural networks or pre-trained models to make predictions. We also compare different models for the action classification task with our dataset. The highest result is on the CNN+LSTM model, achieving 75.88% accuracy on the BCiPS dataset.

Keywords: Video Classification· BCiPS · Pre-school · Deep Learning · Surveillance Cameras

1 INTRODUCTION

The purpose of action classification in videos is to determine what is happening in the video. Human Activities can be classified into various categories including Human-Human interaction and Human-Object interaction. This classification is based on human actions specified by their gestures, poses,... Human action recognition is challenging due to variations in motion, illumination, partial occlusion of humans, viewpoint, and anthropometry of people involved in the various interactions[2]. Additionally, the issue of a person's style when performing a gesture, not only in terms of timing but also in the way the gesture is performed.

In recent years, advancements in neural network technology and deep learning have become effective methods for many tasks, including action classification in videos. Along with the emergence of pre-trained models, the task of action classification has also significantly improved. However, using videos from cameras in schools to classify videos is still a new data domain and there are not many studies on this domain.

Therefore, in this study, we performed the task of classification of videos based on the behaviors of children in preschool through surveillance cameras. The input to the task as a video and the output as the classification of normal or abnormal human actions. After completing the experiments and research, we achieved two results:

- We defined the problem and provided guidelines for creating the BCiPS dataset, which consists of 4,268 videos ranging from 3 to 6 seconds long. BCiPS is one of the first datasets containing videos extracted from surveillance cameras in the preschool domain.

- We experimented with deep learning models for the video classification problem on the BCiPS dataset. Furthermore, we evaluated the performance of the models. The best-performing model was CNN+LSTM, achieving an accuracy of 75.88%. Additionally, we analyzed error cases to identify challenging scenarios that could help future research avoid similar errors and improve the model's performance for real-world applications.

The remaining parts of this paper are as follows. Related work is the section that introduces datasets and methods for classifying human actions in videos. Dataset Introduction will present the steps to create a complete dataset. Methods will show the models used in the study. Experiments and Results will provide the procedure and results achieved in the experiment. Finally, the Conclusion and Future Directions summarize the research process and propose new directions for the task.

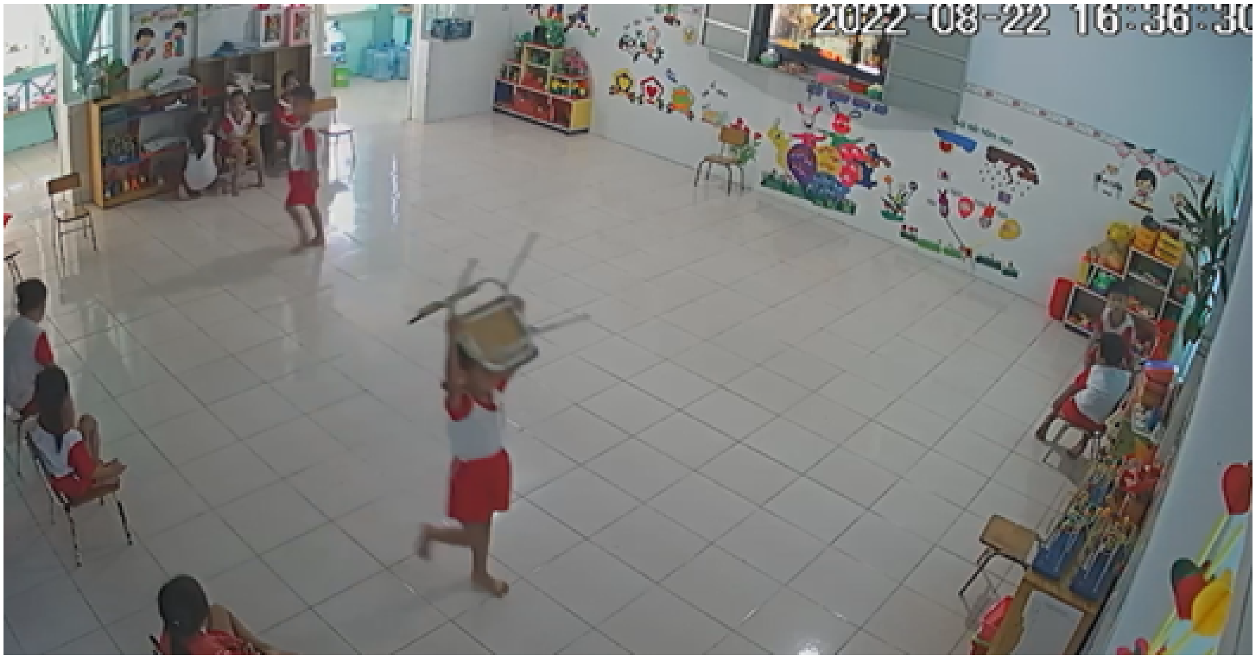

Fig. 1: A frame of a video input. The output of this video is the label "Abnormal" because of the action of a child carrying a heavy object (a chair).

2 RELATED WORKS

2.1 Related Methods

Currently, there are many approaches to solving the human action recognition (HAR) task in videos, including both machine learning and deep learning. Many models have been created based on both for the HAR task, and the existence of these models has made the results of action classification on video more compelling.

Machine learning: In the human action recognition task (HAR) on videos, machine learning has played a significant role in creating models to solve the HAR task. Many models have been created for this task, such as KNN, SVM, Random Forest, and Logistic Regression... These models have been used to solve the HAR task in many papers [[22], [8], [1]].

Deep learning: Long short-term memory (LSTM) is a structure similar to RNN, proposed in 1997 by two scientists, Sepp Hochreiter and Jurgen Schmidhuber (ref). LSTM has feedback connections and can handle single data points such as images, sounds, and texts. Some HAR tasks have also used LSTM for experiments that have been published ([13]). 2D-CNN is a 2D convolutional neural network; some papers have been issued with this method [[15], [16]]. 3D-CNN: the whole video is converted into an input tensor, with the height and width of the video representing spatial while the number of frames representing temporal. The model will use 3D-CNN to process the 3D input and return the classification results of the video. Many papers using this method have been published [[2], [21], [6]]. Some models use this method for human action classification in videos, such as MoViNets, EfficientNets, and I3D, ...

2.2 Related Datasets

Many datasets have been published for human action recognition in videos. These are large-scale datasets with many labels and various topics in various fields. Some examples of these datasets include HMDB-51: This dataset contains 51 action classes and 6,766 clips extracted from 3,312 videos. The UCF101 [19] dataset contains 101 action classes, grouped into five types of actions, and a total of 13,320 clips extracted from 2,500 videos. Kinetics Human Action Video Dataset [9] has 400 action classes and 306,245 clips extracted from 306,245 videos. NTU RGB+D [17]: This human action recognition dataset in a school domain. It includes 60 action classes, divided into three groups: 40 daily actions, nine health-related actions, and 11 mutual actions. Kinetics-700 [4]: This dataset contains over 650,000 videos and 700 action classes, with an average length of 10 seconds per video. The data domain of the dataset includes sports actions, movies, and online videos,... The dataset was collected from YouTube, sports, and movie websites. The something-Something dataset [7] contains over 1000,000 videos with 174 action classes.

We realize that there are few datasets that exist in the school domain, especially in preschool. Due to the issue of personal privacy, child protection is a sensitive issue, affecting many aspects of the lives of those recorded by the camera. Understanding such a situation, we have created the BCiPS dataset, including videos on the preschool domain for our research. Regarding legal matters, we have contacted relevant competent parties and obtained permission to provide videos from the preschool surveillance camera.

3 THE BCiPS DATASET

3.1 Data Collection and Pre-Processing

Data collection: we describe the process of creating the BCiPS dataset for the task of human action recognition in videos. The dataset was created by collecting video footage captured through two surveillance cameras located in two classrooms of a preschool. The cameras recorded the daily activities of students during various times of the day, including studying, entertaining, and napping.

Data pre-processing: video data was collected in .dav format and then converted to .mp4 format using the web tool 123APPS [1]. This format is widely used and easy to work with using various tools. As the focus of the study was on human actions, we removed any periods of time where there were no students in the classroom from the videos. Subsequently, we used the FFmpeg library provided by Python to split the videos into segments of 3-6 seconds. Finally, labeling was performed. After the data processing step, we obtained a dataset consisting of 4,268 videos divided into 2 labels, with a total size of 19.9GB and a total duration of 21,353 seconds.

3.2 Guidelines

The purpose of the task is to classify whether the input video contains normal or abnormal actions. Therefore, the output will be one of two labels: "Normal" or "Abnormal". We define the labels as follows:

[1] https://123apps.com/

- **Normal:** videos containing actions that regularly occur and are not dangerous in the observed environment. For example, walking, talking, eating, playing, studying, ..., performed normally and regularly in a school or preschool domain.
- **Abnormal:** videos contain unusual and dangerous actions for children, such as fighting, falling, chasing, carrying heavy or big objects, and other abnormal activities. These actions can be dangerous, leading to injury and accidents or affecting the mental health of the children.

For videos where children are not sleeping, we propose watching the actions in the video and evaluating whether they fall within the range of normal actions. If the activities in the video do not belong to the list of abnormal actions and do not harm children, we will label the video as "Normal".

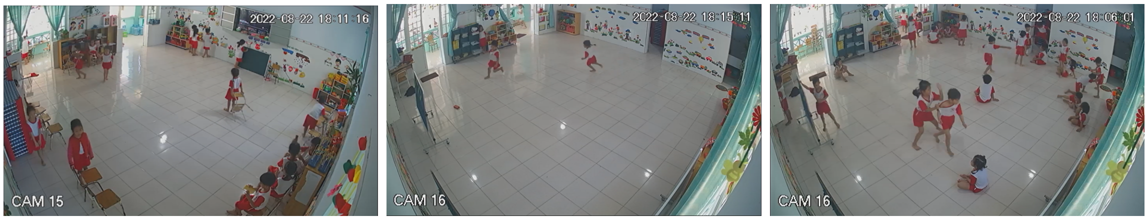

Fig. 2: Three frame of three videos labeled as 'Abnormal'.

Regarding the frame captured from CAM15, it shows a child carrying a heavy object (a chair) while another child is standing on the chair. This could potentially lead to a fall and pose danger to the children. As for the middle frame (left panel of CAM16), there are two children chasing each other, which could also lead to falling and cause danger. In the frame on the right panel of CAM16, a girl is seen pulling a shirt and physically impacting someone else in the frame. Therefore, three videos is labled as "Abnormal".

3.3 Annotators Agreement

We conducted the agreement among labeling annotators to reach consensus and objectivity before building the dataset. Five annotators independently labeled 150 identical videos. We then calculated the annotation agreement between each pair of annotators using Cohen's Kappa index and the average annotations agreement [5]. According to the ranking table for annotation agreement in categorical data [12], Table 1 shows that the average agreement among pairs was 0.68, which is considered good as it falls between the range of 0.6 and 0.81.

Table 1: The agreement among annotators as measured by Cohen's Kappa.

	A1	A2	A3	A4	A5
A1	1	0.70	0.64	0.71	0.73
A2	-	1	0.46	0.56	0.56
A3	-	-	1	0.81	0.73
A4	-	-	-	1	0.86
A5	-	-	-	-	1
Average	0.68				

In addition to calculating the annotation agreement among each pair using Cohen's Kappa index, we also calculated the agreement among all five annotators using Krippendorff's alpha [11]. The agreement was 67.55%, which is higher than the minimum acceptable value of 66.7% for Krippendorff's alpha. Therefore, our agreement after labeling the data according to the guidelines in subsection 3.2 was higher than the minimum acceptable result for Krippendorff's alpha and achieved good agreement according to Cohen's Kappa index. As a result, we can officially label the collected data.

During the labeling process, it is inevitable to encounter challenging and ambiguous cases. To address these issues, the collaborators will have meetings to discuss and agree on how to label these difficult and ambiguous cases and update the guidelines to make them more complete.

3.4 Dataset Analysis

After the data collection, preprocessing, and labeling steps, BCiPS include 2 labels, 4,268 videos with a total size of 19.9GB, and a total length of 21,353 seconds. We divided the dataset into three sets, training, development, and testing, with a ratio of 8:1:1. The table 2 provides an overview of the dataset.

Figure 3 and table 2 show that the number of Normal and Abnormal labels is relatively balanced. In three datasets, the ratio of 2 labels is equal. The train set is much larger than the test and validation set because we want the model to learn more cases, covering the data. For the test set and the validation set with a ratio of 1:1, this helps evaluate the most objective and close to reality.

Table 2: Dataset Statistics

	Number of videos			
	Train	Validation	Test	Entire
Entire	3,414	427	427	4,269
Normal	1,760	220	220	2,200
Abnormal	1,654	207	207	2,068

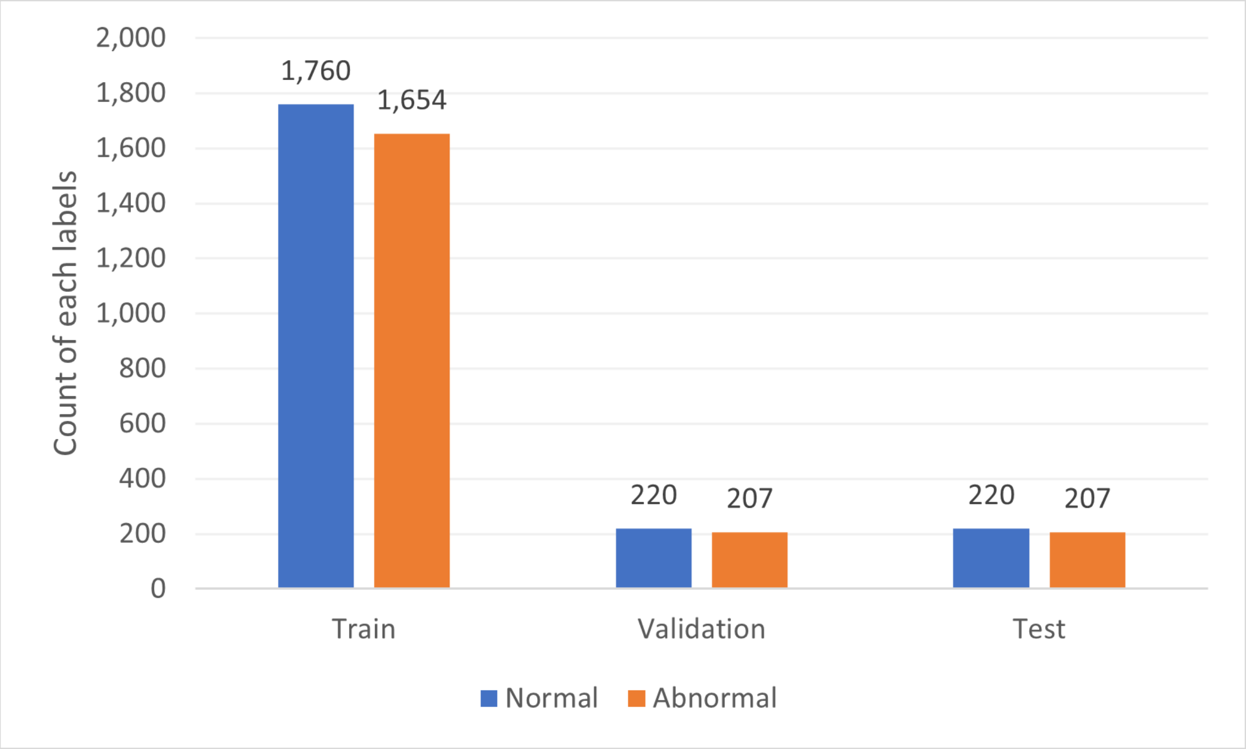

Fig. 3: Distribution of labels in the training, validation, and testing sets.

4 Methodologies

4.1 TimeSformer

As TimeSformer (Time-Space Transformer) [3] is one of the most advanced methods for video classification that has emerged recently, we have decided to use this architecture in our benchmark. The TimeSformer model is designed for videos and is pre-trained on the ImageNet dataset and fine-tuned on our dataset. TimeSformer does not contain convolutional layers, either, and it consists of a self-attention mechanism. TimeSformer adapts transformer architecture for computer vision and video processing tasks.

4.2 MoViNets

MoViNets [10] is a video classification model used for online video streaming or real-time inference in tasks such as action recognition. The classifier is based on efficient and simple 2D frame-level processing to run over the entire video or stream each frame one by one. As they cannot take into account temporal context, they have limited accuracy and may give inconsistent output results from one frame to another. A simple 3D CNN that uses a two-dimensional temporal context can increase accuracy and temporal consistency. These networks may require more resources and since they are oriented toward the future, they cannot be used for data transmission.

4.3 (2+1)D Resnet-18

The following 3D convolutional neural network model is based on what D. Tran et al. published. The $(2 + 1)$D convolution allows for the separation of spatial and temporal dimensions, thus creating two separate steps. One advantage of this method is that analyzing combinations into spatial and temporal dimensions helps save parameters.

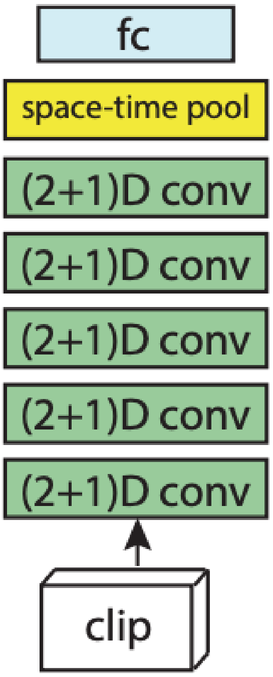

Fig. 4: Architecutre of (2+1)D Convolution.

4.4 EfficientNetB0

EfficientNet [20] was introduced by Tan and Le, who studied the scaling of models and determined that carefully balancing the depth, width, and resolution of a network can lead to better performance. They proposed a novel scaling method that evenly scales all dimensions of a network, including depth, width, and resolution. They used a neural architecture search tool to design a new base network and extended it to obtain a family of deep learning models. There are 8 variants of EfficientNet (B0 through B7), with EfficientNetB0 having 5.3 million parameters. Figure 5 illustrates the architecture of the EfficientNet network.

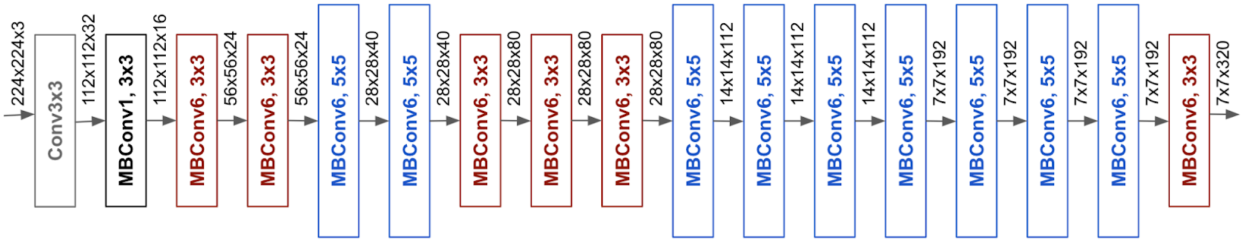

Fig. 5: Architecture of EfficentNet.

4.5 CNN+LSTM (ConvLSTM2D)

The Convolutional LSTM (ConvLSTM) network was first introduced in the work of [18]. In a fully connected LSTM network, flattening the image into a 1D space does not retain any spatial information, hence the need for CNN to extract spatial features and transform them into a 1D vector space. Therefore, the ConvLSTM network was proposed for video classification tasks [14], using 2D structures as inputs. It can directly work with a sequence of images and perform convolutional operations on the input images to extract spatial features, while LSTM layers can extract temporal dynamics between frames. Therefore, the ConvLSTM network can essentially capture both spatial and temporal signals, which cannot be achieved by fully connected LSTM.

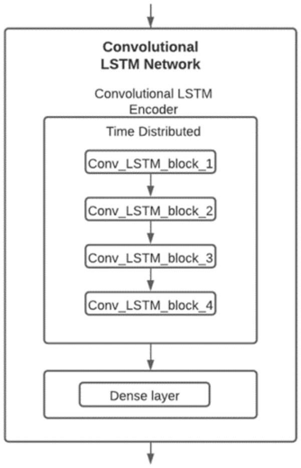

Fig. 6: Architecture of CNN-LSTM

4.6 CNN+SVM

CNN is used to extract features from video data. CNN is typically used to learn 2D image features from each frame. The output results of CNN for each frame generates a feature vector. It is then used to train SVM model.

4.7 CNN+Random Forest

CNN is used to extract features from video data. CNN is typically used to learn 2D image features from each frame. The output results of CNN for each frame generates a feature vector. It is then used to train Random forest model.

5 EXPERIMENT AND RESULTS

5.1 Technical aspects of the experiment

In this section, we present 5 steps to obtain the results of this research:

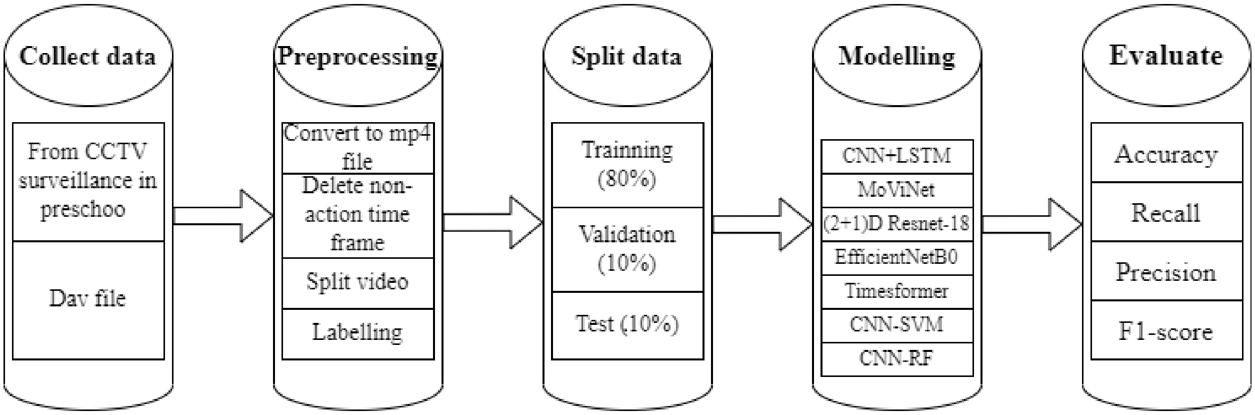

Fig. 7: Experimental Procedure

Step 1: Collect data from a preschool using .dav files, which are videos containing the actions of children recorded from two surveillance cameras placed in the classroom.

Step 2: Preprocess the data.

- Convert the videos to .mp4 format.
- Remove time periods without students, as these can make learning difficult for the model.
- From the remaining videos, we extract clips that are 3-6 seconds in length.
- Label the extracted videos. There are two labels: normal and abnormal.

Step 3: Split the dataset of 4.268 videos into training, validation, and testing sets with an 8:1:1 ratio.

Step 4: Implement 7 models: CNN-LSTM, MoViNet, (2+1)D Resnet-18, EfficientNetB0, Timesformer, CNN-SVM, CNN-RandomForest. Table 3: The hyperparameters corresponding to each model is in Table 3:

Step 5: Evaluate the models using accuracy, recall, precision, and F1-score metrics.

Table 3: Hyper-parameters of Models

Model	FPS	Batch size	Learning rate	Epoch
CNN + LSTM	10	64	–	50
CNN + SVM	10	–	–	1
CNN + RandomForest	10	–	–	1
Movinet	10	8	1.00e-05	3
(2+1) Resnet-18	10	8	1.00e-05	3
EfficientNetB0	8	8	1.00e-05	5
Timesformer	8	4	1.00e-05	3

5.2 Results

Our task is a classification task. The evaluation metrics used by our team to assess the model include accuracy, recall, precision, and F1-score. After conducting experiments on 427 videos from the test set, we obtained the results as presented in Table 4:

Table 4: Models Performance

	Accuracy	Recall	Precision	F1
CNN + LSTM	**75.88**	**75.63**	**75.43**	**75.53**
Timesformer	74.71	74.80	74.82	74.81
CNN + RandomForest	74.00	73.97	73.98	73.98
EfficientNetB0	73.30	73.48	73.70	73,59
Movinet	70.02	69.61	71.30	70.44
CNN + SVM	65.57	65.61	65.59	65.60
(2+1)D Resnet-18	61.12	60.42	63.37	61,86

Table 4 shows that the CNN model combined with LSTM achieved the highest results in all four metrics: accuracy, recall, precision, and F1-score with 75,88%, 75,63%, 75,43%, 75,62%, respectively. The results of the CNN-LSTM model were significantly different from the other models. Besides, the (2+1)D Resnet-18 model had the lowest performance with only 61,12%, 60,42%. 63,37%, 61,86%, much lower than the other models. It can be seen that with our dataset, the pre-trained models resulted in lower performance than those combining traditional models.

5.3 Error analysis

Table 5 show that the CNN+LSTM model achieved the best performance in terms of accuracy and recall (we included recall as a performance metric to minimize false negatives). Therefore, we conducted a detailed error analysis to identify the specific challenges faced by the model.

Table 5: Recall and Precision Scores per Label for the CNN+LSTM Model.

	Normal	Abnormal
Recall	83.64%	67.63%
Precision	73.31%	79.55%

Based on the performance metrics in the table 5 and figure 8, it can be seen that the "Normal" class has a precision of 73.31%, while the "Abnormal" class has a precision of 79.55%. This suggests that the model tends to classify videos as abnormal rather than ignoring abnormal videos. The difference in precision between the two classes indicates that the model is moving in the right direction for the development of the dataset but still requires further improvement. The error rate and the difference in the precision of the two classes suggest that the model is still prone to misclassifications

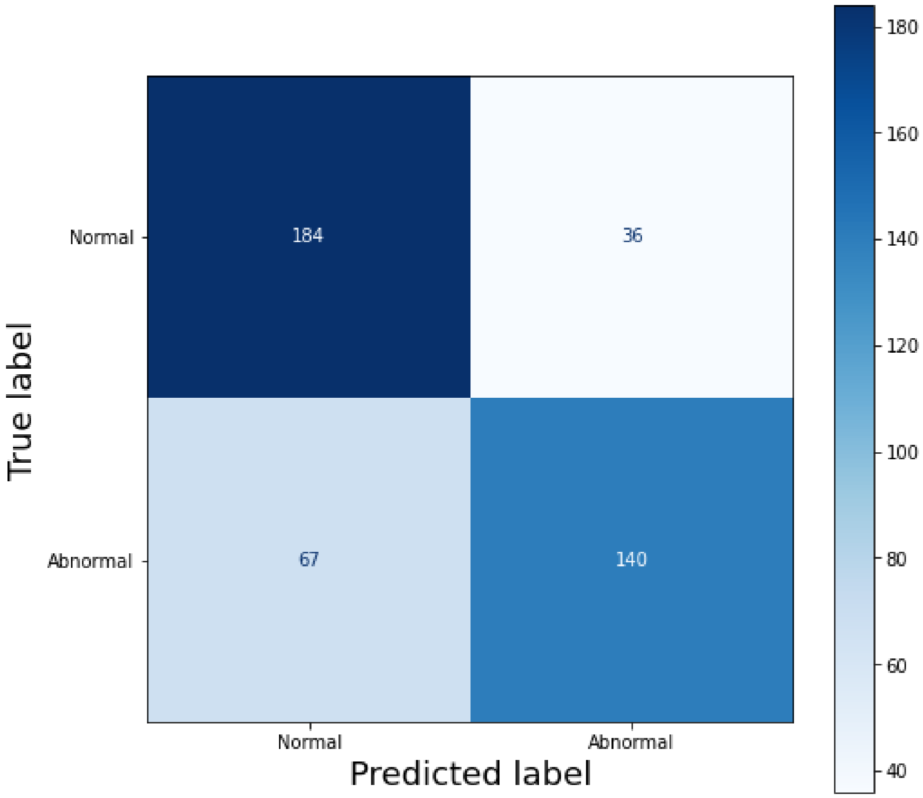

Fig. 8: Confusion Matrix

and needs to be fine-tuned for better accuracy.

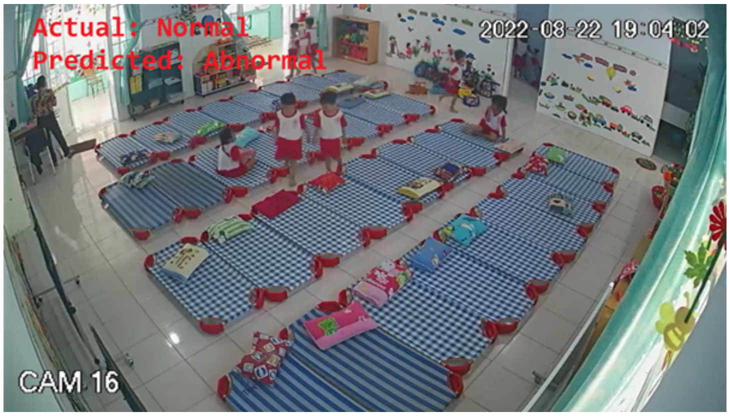

Fig. 9: Predicted label and actual label of a sample video from the test set.

Figure 9, we can see that the video contains many ambiguous actions (e.g., "Running"), which caused the model to predict it as abnormal. The main reason is that some abnormal behaviors have not been fully defined, the video also contains some ambiguous actions, and the dataset needs to be more diverse.

6 Conclusion-Future Works

In this paper, we have created a dataset for classifying the behaviors of children in preschool named BCiPS. There are two labels, "Normal" and "Abnormal". After experimenting with the dataset on several video classification models, including basic, combined, and pre-trained models, we achieved the highest accuracy of 75.88% with the CNN + LSTM model. This demonstrates the potential of the dataset we have constructed for video classification in a preschool domain. However, specific errors still need to be addressed, which hinder the optimal performance of the models. In order to achieve real-world application, we plan to improve the dataset by collecting more data, re-labeling the data more effectively, and establishing stricter guidelines for data labeling.

Observing the existing limitations in the dataset, we intend to take some improvement steps towards applying the dataset in practical applications. Specifically, we will collect additional data, re-label the dataset more efficiently, and establish stricter labeling guidelines. Finally, we will explore suitable models for this dataset.

REFERENCES

1. Ahmad, T., Wu, J., Khan, I., Rahim, A., Khan, A.: Human action recognition in video sequence using logistic regression by features fusion approach based on cnn features. International Journal of Advanced Computer Science and Applications (11) (2021)
2. Alfaifi, R., Artoli, A.M.: Human action prediction with 3d-cnn. SN Computer Science **1**, 1–15 (2020)
3. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
4. Carreira, J., Noland, E., Hillier, C., Zisserman, A.: A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987 (2019)
5. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
6. Diba, A., Fayyaz, M., Sharma, V., Karami, A.H., Arzani, M.M., Yousefzadeh, R., Van Gool, L.: Temporal 3d convnets: New architecture and transfer learning for video classification. arXiv preprint arXiv:1711.08200 (2017)
7. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
8. Hbali, Y., Hbali, S., Ballihi, L., Sadgal, M.: Skeleton-based human activity recognition for elderly monitoring systems. IET Computer Vision **12**(1), 16–26 (2018)
9. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
10. Kondratyuk, D., Yuan, L., Li, Y., Zhang, L., Tan, M., Brown, M., Gong, B.: Movinets: Mobile video networks for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16020–16030 (2021)
11. Krippendorff, K.: Content analysis (2004)
12. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)
13. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. pp. 816–833. Springer (2016)
14. Luo, W., Liu, W., Gao, S.: Remembering history with convolutional lstm for anomaly detection. In: 2017 IEEE International Conference on Multimedia and Expo (ICME). pp. 439–444. IEEE (2017)

15. Núñez-Marcos, A., Azkune, G., Arganda-Carreras, I.: Vision-based fall detection with convolutional neural networks. Wireless communications and mobile computing **2017** (2017)

16. Pham, H.H., Khoudour, L., Crouzil, A., Zegers, P., Velastin, S.A.: Video-based human action recognition using deep learning: a review. arXiv preprint arXiv:2208.03775 (2022)

17. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)

18. Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems **28** (2015)

19. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

20. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)

21. Uchiyama, T., Sogi, N., Niinuma, K., Fukui, K.: Visually explaining 3d-cnn predictions for video classification with an adaptive occlusion sensitivity analysis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1513–1522 (2023)

22. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 588–595 (2014)