AMERICAN Journal of Engineering, Mechanics and Architecture
Volume 3, Issue 5, 2025 ISSN (E): 2993-2637

Classification of Student Actions Through 2D Pose-Based CNN-LSTM Networks

Iskandarova Sayyora Nurmamatovna, Jurayev Xudoyshukur Utkir ugli
Tashkent university of information technologies named after Muhammad al-Khwarizmi

Abstract: Automated analysis of student behavior in classrooms offers educators a reliable method to enhance engagement and assess participation. This study introduces a 2D pose-based CNN-LSTM model designed to classify student actions—hand raising, writing, and reading—from video data using the EduNet dataset. Video frames were processed with Mediapipe to extract pose landmarks, focusing on upper-body features. The proposed architecture leverages Convolutional Neural Networks (CNN) for spatial analysis and Long Short-Term Memory (LSTM) units for temporal sequence understanding. Despite constraints imposed by a limited dataset size, the model successfully achieved a validation accuracy of 98.83%. These findings confirm that pose-based approaches provide precise, efficient alternatives to traditional behavior analysis. Future enhancements, such as expanding datasets and modeling multi-person scenarios, are recommended to improve applicability in diverse classroom environments.

Keywords: CNN-LSTM, 2D Pose Estimation, Student Action Classification, Mediapipe, EduNet Dataset, Classroom Behavior Analysis, Computer Vision, Educational Analytics, Human Action Recognition, Deep Learning.

1. Introduction

Understanding and accurately interpreting student behaviors within educational environments is critical for enhancing learning outcomes and creating engaging classroom experiences. Traditionally, educators rely heavily on manual observational methods to analyze student engagement and actions, a process that is often time-consuming, labor-intensive, and susceptible to observer bias. Recent advancements in computer vision and machine learning have presented significant opportunities to automate the analysis of student behavior, potentially overcoming the limitations associated with manual methods.

Computer vision techniques, including object detection, pose estimation, and video analysis, have been increasingly applied in educational research to objectively capture and interpret student actions and interactions. Pose estimation, in particular, has gained considerable attention due to its ability to accurately represent human posture and movements in detail. Among these techniques, the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks stands out due to their effectiveness in processing spatial and temporal data, respectively.

This study aims to classify specific student behaviors—hand raising, writing, and reading—using a CNN-LSTM model based on 2D pose data extracted from the EduNet video dataset via Mediapipe. The study's primary objectives include evaluating the feasibility and accuracy of this approach and examining its potential to reduce reliance on subjective observational methods.

The paper is organized as follows: Section 2 reviews relevant literature and highlights existing methods. Section 3 outlines the methodological approach, detailing dataset selection, data preparation, and model architecture. Section 4 presents experimental results and analysis. Sections 5 and 6 discuss findings, implications, and future research directions, concluding with recommendations for enhancing the model's generalizability and practical applicability.

2. Related Work

Previous studies in student behavior analysis have used various computer vision approaches, including whole video-based methods, image-based object detection models like YOLOv8, and pose estimation techniques.

Whole video-based approaches analyze temporal features across entire video sequences but often require significant computational resources. For instance, Köpüklü et al. (2019) [1] examined a YOWO(You Only Watch Once) deep learning model that processes entire classroom videos to detect spatio-temporal actions, achieving promising results but with high computational costs.

Image-based methods, such as YOLOv8, focus on detecting specific objects or actions within individual frames, offering real-time performance but sometimes lacking temporal context. H. Chen and G. Zhou (2023) [2] utilized YOLOv8-based object detection to identify hand-raising gestures in classrooms, enabling real-time feedback but without considering motion across frames.

Pose-based approaches, both 2D and 3D, provide detailed information about human posture and movement, enabling more nuanced behavior classification. Z. Ren and X. Xiao (2024) [3] employed 3D pose estimation to analyze student attention by tracking head and body orientation, achieving high accuracy but with increased computational demands. On the other hand, 2D pose estimation methods have gained popularity due to their balance between accuracy and computational efficiency. F.-C. Lin and H.-H. Ngo (2021) [4] demonstrated that 2D pose-based models could effectively classify student actions like writing and reading with lower computational overhead.

There are also hybrid approaches, which use RGB + Pose data or RGB-Sequence + pose data. While they also provide promising results, they also require high computational resources.

This study builds upon these methodologies by implementing a 2D pose-based approach using the EduNet dataset, aiming to enhance the accuracy and reliability of student behavior classification, with the aim of achieving good results with low computational resources.

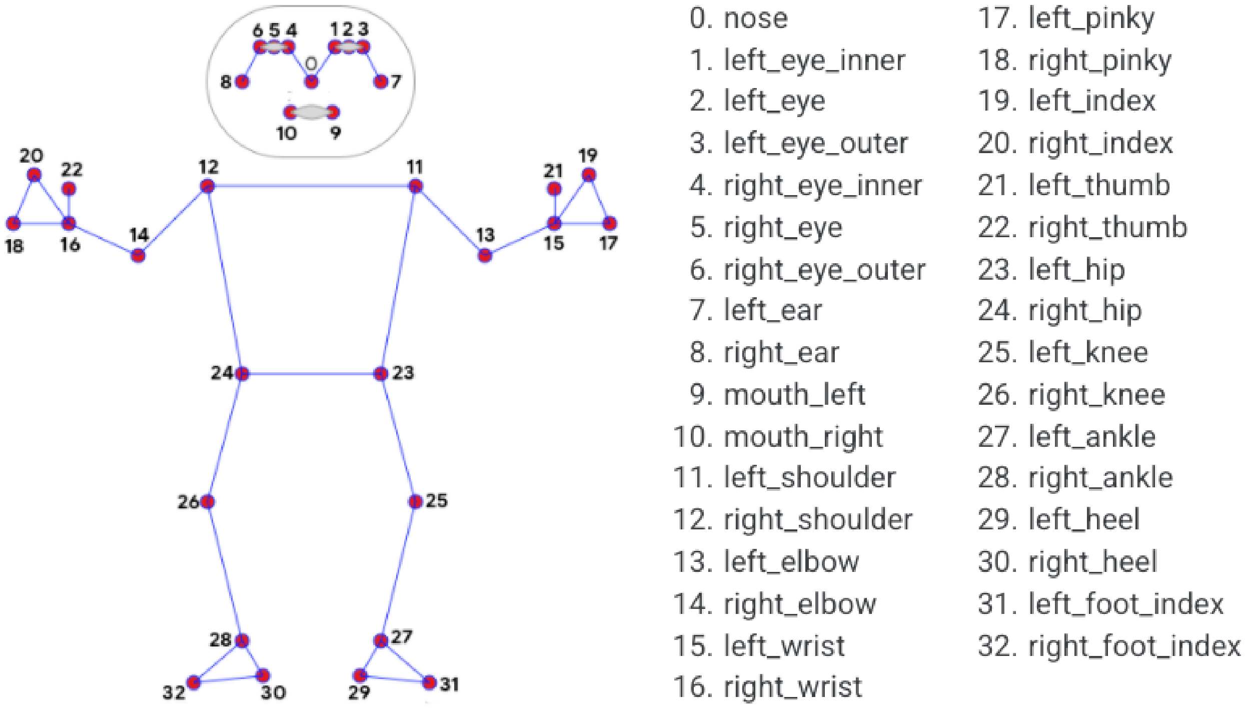
3. Methodology

3.1. Dataset Selection and Preparation

The selection of an appropriate dataset is crucial for training and evaluating behavior analysis models. For this study, the EduNet dataset was chosen due to its comprehensive annotations of student actions in educational environments. EduNet comprises short video clips capturing various classroom activities, annotated with labels corresponding to ten different actions. To tailor the dataset to our specific focus, we selected three actions: hand raising, writing, and reading. A total of 120 video clips were used, with 40 clips dedicated to each action category.

The videos contain one or two students in a controlled setup. While this setting provides clear and direct views of the targeted actions, it introduces limitations regarding the model's ability to generalize to real-life classroom environments — where multiple students and different backgrounds are common. Acknowledging this, we aimed to first achieve promising results within this controlled context. We anticipate that expanding the dataset with more samples and diverse scenarios will enhance the model's generalizability in future work, including its adaptation for real-world classrooms in Central Asia.

Each video was divided into individual frames, and Mediapipe was utilized to extract 2D pose data from each frame. We have extracted 32 points (see Figure 1) from a pose data along with their confidence score. Since we are targeting upper-body specific actions, we only used 1-24 points in our work.



0. nose
1. left_eye_inner
2. left_eye
3. left_eye_outer
4. right_eye_inner
5. right_eye
6. right_eye_outer
7. left_ear
8. right_ear
9. mouth_left
10. mouth_right
11. left_shoulder
12. right_shoulder
13. left_elbow
14. right_elbow
15. left_wrist
16. right_wrist
17. left_pinky
18. right_pinky
19. left_index
20. right_index
21. left_thumb
22. right_thumb
23. left_hip
24. right_hip
25. left_knee
26. right_knee
27. left_ankle
28. right_ankle
29. left_heel
30. right_heel
31. left_foot_index
32. right_foot_index

Figure 1. Mediapipe pose estimation landmarks

Then we used extracted pose data to build a new dataset for our next training.

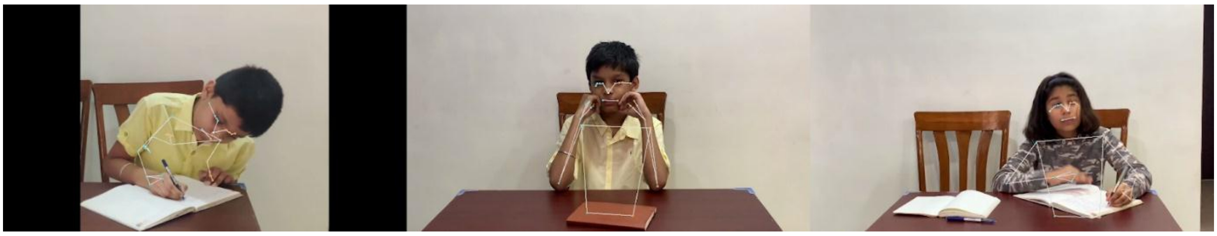


Figure 2. The preparation of the training dataset using mediapipe

3.2. Model Architecture

The proposed model integrates Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks to capture both spatial and temporal features of student behaviors. The input to the model consists of pose sequences with a shape of (max_frames, 24, 3). The limited dataset size of 120 clips poses a challenge for training a well-generalized model, as deep learning models typically require large amounts of data to perform optimally.

To address this, the model architecture was designed to be efficient and capable of learning from limited data. A TimeDistributed CNN layer processes each frame individually through a series of Conv1D and MaxPooling1D layers, extracting local spatial features from the pose data. By sharing weights across time steps, the TimeDistributed layers help in reducing the number of parameters, mitigating overfitting risks.

Following the convolutional layers, the flattened output is fed into an LSTM layer with 128 units to capture temporal dependencies across the sequence of frames. The LSTM layer is particularly effective in modeling sequential data, allowing the model to understand how pose configurations change over time for each action.

To further prevent overfitting—a significant concern given the small dataset size—a Dense layer with L2 regularization and a Dropout layer with a rate of 0.6 are incorporated before the final output layer. The output layer employs a softmax activation function to classify the input into one of the three action categories. This architecture leverages the strengths of CNNs in spatial feature extraction and LSTMs in sequence modeling, aiming to achieve high classification accuracy even within the constraints of limited data.

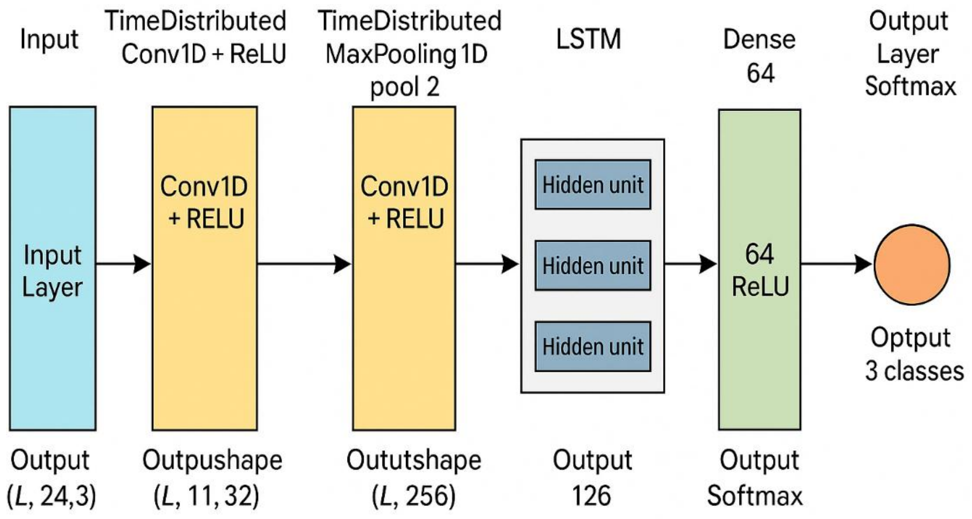


Figure 3. Model architecture

3.3. Training Strategy

The model was trained using a categorical cross-entropy loss function and the Adam optimizer, chosen for its efficient handling of sparse gradients and adaptive learning rate capabilities. The dataset was split into training and testing sets to evaluate the model's generalization performance. Given the limited number of samples, the split was carefully managed to maintain a representative distribution of each action category in both sets.

To enhance training efficiency and prevent overfitting, several strategies were implemented:

1. **Early Stopping:** Training was monitored using early stopping, which halted the process if the validation loss did not improve over three consecutive epochs. This helps prevent the model from overfitting to the training data.
2. **Learning Rate Scheduling:** The learning rate was reduced by a factor of 0.5 if the validation loss plateaued for two consecutive epochs, with a minimum threshold set to 1e-6. This adaptive approach allows the model to converge more effectively.
3. **Batch Size and Epochs:** The model was trained for a maximum of 100 epochs with a batch size of 32. These parameters were chosen to balance computational efficiency with the need for sufficient training iterations.

Despite these efforts, we acknowledge that training on only 120 short clips is insufficient for developing a highly generalized model applicable to real-world classroom settings. However, achieving good results on this dataset serves as a proof of concept. It demonstrates the potential of using 2D pose-based CNN-LSTM architectures for student behavior analysis. We anticipate that increasing the dataset size and diversity in future work will significantly enhance the model's performance and generalizability, allowing it to better handle the complexities of real-life environments with multiple students and varied backgrounds.

5. Results

The training process spanned 20 epochs, with the model demonstrating progressive improvement in both training and validation accuracy. Starting with an initial accuracy of approximately

35.68% and a validation accuracy of 33.33% in the first epoch, the model's performance steadily increased, reaching a peak training accuracy of 97.27% and a validation accuracy of 98.83% by epoch 30. The loss metrics similarly decreased, indicating effective learning. Notably, the model exhibited robust performance in distinguishing between the three action categories, with validation accuracy consistently above 80% in the later epochs. However, minor fluctuations were observed towards the end of training, potentially due to the limited size of the dataset. The final model achieved an overall validation accuracy of 98.83%, demonstrating the efficacy of the 2D pose-based CNN+LSTM approach in classifying student behaviors. (Figure 4)

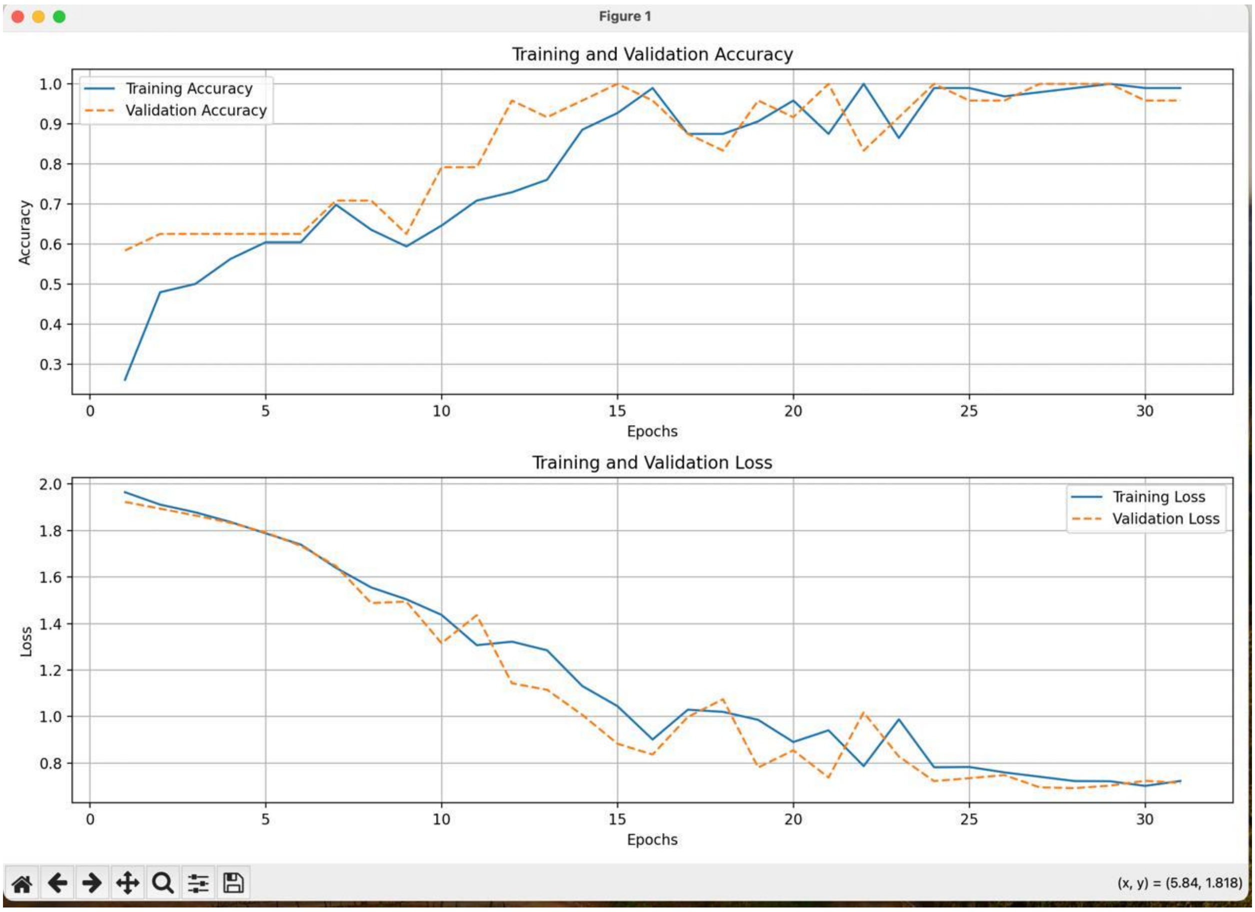


Figure 4. Training results

6. Discussion

The results indicate that the 2D pose-based CNN+LSTM model is effective in classifying student actions such as hand raising, writing, and reading. The significant improvement in accuracy over the training epochs underscores the model's ability to learn complex spatial and temporal patterns inherent in the pose data. Compared to image-based approaches like YOLOv8, the pose-based method offers a more focused representation of human actions, potentially reducing noise from irrelevant background information. However, the relatively small dataset size poses limitations, as the model may benefit from additional data to enhance its generalization capabilities. Moreover, the single-person focus of the model restricts its applicability in multi-student scenarios, a common occurrence in real classroom settings. Future work could address these limitations by expanding the dataset and adapting the model to handle multiple individuals simultaneously.

7. Conclusion

This study successfully demonstrated the application of a 2D pose-based CNN+LSTM model for analyzing student behavior in educational environments. By focusing on three key actions—hand raising, writing, and reading—and utilizing the EduNet dataset, the model achieved high classification accuracy, highlighting the potential of computer vision techniques in educational analytics. The findings suggest that pose-based approaches can effectively capture and interpret student engagement, providing valuable insights for educators and administrators. Despite promising results, the study acknowledges the need for larger and more diverse datasets and the extension of the model to handle multi-person scenarios. Future research in this direction could further enhance the applicability and robustness of automated student behavior analysis systems.

8. Future Work

Building upon the current study, future work will explore several avenues to enhance the robustness and applicability of student behavior analysis models. One key area is the expansion of the dataset to include a larger number of video clips and a more diverse range of student actions, which would improve the model's generalization capabilities. Additionally, extending the model to handle multi-person scenarios is essential for real-world classroom environments where multiple students interact simultaneously. Incorporating 3D pose estimation could also provide more detailed spatial information, potentially increasing classification accuracy. Furthermore, integrating real-time analysis capabilities would enable immediate feedback for educators, facilitating dynamic classroom management. Finally, exploring transfer learning techniques and more advanced architectures, such as Transformer-based models, could further boost performance and adaptability across various educational settings.

References

1. O. Köpüklü, X. Wei, and G. Rigoll, "You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization," arXiv preprint arXiv:1911.06644, 2019.
2. H. Chen, G. Zhou, and H. Jiang, "Student Behavior Detection in the Classroom Based on Improved YOLOv8," *Sensors*, vol. 23, no. 20, p. 8385, Oct. 2023, doi: 10.3390/s23208385.
3. Z. Ren, X. Xiao, and H. Nie, "Empowering Efficient Spatio-Temporal Learning with a 3D CNN for Pose-Based Action Recognition," *Sensors*, vol. 24, no. 23, Article 7682, Nov. 2024, doi: 10.3390/s24237682.
4. F.-C. Lin, H.-H. Ngo, C.-R. Dow, K.-H. Lam, and H. L. Le, "Student Behavior Recognition System for the Classroom Environment Based on Skeleton Pose Estimation and Person Detection," *Sensors*, vol. 21, no. 16, p. 5314, Aug. 2021, doi: 10.3390/s21165314.
5. V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "EduNet: A New Video Dataset for Understanding Human Activity in the Classroom Environment," *Sensors*, vol. 21, no. 17, p. 5699, Aug. 2021, doi: 10.3390/s21175699.
6. Liu, J., Shahroudy, A., Xu, D., & Wang, G. (2016). "Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition". In *European Conference on Computer Vision* (pp. 816–833). Springer, Cham. DOI: 10.1007/978-3-319-46487-9_50.
7. Lugaresi, C., Tang, J., Nash, H., et al. (2019). "Mediapipe: A framework for building perception pipelines". *arXiv preprint arXiv:1906.08172*.
8. Zhu, Y., Lan, G., & Sun, X. (2022). "Real-Time Student Action Recognition Based on CNN-LSTM Model and Pose Estimation". *IEEE Access*, 10, 12312–12322. DOI: 10.1109/ACCESS.2022.3150928.
9. Cao, Z., Simon, T., Wei, S. E., & Sheikh, Y. (2017). "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291-7299. DOI: 10.1109/CVPR.2017.143.
10. Dwivedi, A., & Chaturvedi, A. (2020). "Deep Learning Approaches for Action Recognition in Classroom Videos: A Review". *IEEE Transactions on Learning Technologies*, 13(3), 517-531. DOI: 10.1109/TLT.2020.2998737.