Matrix: Jurnal Manajemen Teknologi dan Informatika
Volume 16 Issue 2 Year 2026 Pages 74-85
DOI: http://dx.doi.org/10.31940/matrix.v16i2.74-85
URL: http://ojs2.pnb.ac.id/index.php/matrix

Video-based human activity recognition analysis using YOLOv26

Santi Rahayu [1*], Imam Riadi [2], Andri Pranolo [3]

[1,3] *Informatics Study Program, Universitas Ahmad Dahlan, Indonesia*
[1] *Informatics Engineering Study Program, Universitas Pamulang, Indonesia*
[2] *Information System Study Program, Universitas Ahmad Dahlan, Indonesia*

*****Corresponding Author: *2536083045@webmail.uad.ac.id*

Abstract: Human Activity Recognition (HAR) is a rapidly growing research field in computer vision and has various applications, such as intelligent surveillance systems, sports analysis, healthcare, and human-computer interaction. This study aims to analyze the performance of the YOLOv26 Classification model in recognizing video-based human activities using the UCF101 dataset. This study uses six activity classes, namely JumpingJack, Punch, PushUps, Typing, WalkingWithDog, and WritingOnBoard. The research method is carried out through video frame extraction using a uniform sampling technique, the formation of an image classification dataset, YOLOv26 model training, and evaluation at the frame and video levels. Experiments were conducted using 36 configuration combinations consisting of four YOLOv26 model variants (YOLOv26-n, YOLOv26-s, YOLOv26-m, and YOLOv26-l), three variations in the number of frames (8, 16, and 24 frames), and three variations in the number of epochs (50, 100, and 150 epochs). Video evaluation was conducted using a majority voting approach with accuracy, precision, recall, and F1-score metrics. The results showed that all configurations produced video accuracy above 93%. The best configuration was obtained with the YOLOv26-m model with 16 frames and 50 epochs, which achieved video accuracy of 98.26%, precision of 98.15%, recall of 98.37%, and F1-score of 98.23%. Confusion matrix analysis showed that most predictions were on the main diagonal, indicating the model's ability to distinguish human activities with a low error rate. These results prove that YOLOv26 Classification has excellent performance for video-based human activity recognition and has the potential to be applied to various applications based on automatic human activity analysis.

Keywords: Human activity recognition, YOLOv26 classification, UCF101 dataset, video classification, majority voting, computer vision.

History Article: Submitted 6 June 2026 | Revised 13 June 2026 | Accepted 22 June 2026

How to Cite: S. Rahayu, I. Riadi, and A. Pranolo, "*Video-based human activity recognition analysis using YOLOv26,*" *Matrix: Jurnal Manajemen Teknologi dan Informatika*, vol. 16, no. 2, pp. 74–85, 2026, doi: 10.31940/matrix.v16i2.74-85.

Introduction

Human Activity Recognition (HAR) is one of the fastest-growing research areas in computer vision and has attracted considerable attention due to its wide range of applications in intelligent surveillance systems, healthcare monitoring, sports analysis, human-computer interaction, and smart environments [1], [2], [3]. The ability of computer systems to automatically recognize and classify human activities from video data enables more efficient monitoring, decision-making, and behavioral analysis in real-world scenarios. With the increasing availability of video acquisition devices and surveillance cameras, the development of accurate and efficient HAR systems has become an important research topic.

Video-based HAR remains a challenging task because human activities contain both spatial and temporal information. Spatial information represents the visual appearance of objects and body postures in individual frames, while temporal information describes motion patterns and activity evolution across consecutive frames [2], [4]. In addition, variations in lighting conditions, camera viewpoints, background complexity, occlusion, and motion speed can significantly affect recognition performance [5]. These challenges require robust deep learning models capable of effectively capturing discriminative features from video sequences.

Recent advances in deep learning have significantly improved HAR performance. Convolutional Neural Networks (CNNs) have been widely used to extract spatial features, while temporal dependencies are commonly modeled using architectures such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), Temporal Convolutional Networks (TCN), and Transformers. Jindal et al. combined feature optimization with ResNet101V2 and achieved an accuracy of 96.16% on the UCF101 dataset [6]. Ahmad et al. employed a VGG16-BiGRU framework for HAR and demonstrated the effectiveness of combining spatial and temporal learning [7]. Similarly, Shin et al. utilized a hybrid ConvNeXt-TCN architecture and reported an accuracy of 98.46% on UCF101 [8]. Other studies have explored Selective Kernel Networks [9], graph-based learning [10], and ensemble voting strategies [11] to improve activity recognition performance.

Several recent studies have also investigated advanced spatiotemporal learning approaches. Vrskova et al. demonstrated that 3D Convolutional Neural Networks (3DCNNs) can effectively learn spatial and temporal information simultaneously [2]. Alshaya introduced Neuro-Prismatic Video Models based on UniFormerV2 to capture complex temporal relationships in video data [12]. Zahra et al. proposed Adaptive Text Prompting for zero-shot action recognition [13], while Li et al. introduced Mamba-YOWO for efficient spatiotemporal action detection [4]. Furthermore, Isik and Ozcan proposed a novel activation function that improved HAR classification performance on benchmark datasets [14]. Although these methods have achieved promising results, many of them require complex architectures and high computational resources, making practical deployment more challenging.

The YOLO (You Only Look Once) family has become one of the most successful deep learning frameworks for real-time visual recognition tasks due to its balance between accuracy and computational efficiency. Previous studies have demonstrated the effectiveness of YOLO-based approaches for activity recognition and action localization [15], [16]. More recently, YOLOv26 has been introduced as a new generation architecture featuring an NMS-Free design and end-to-end optimization, which significantly improves inference efficiency and deployment performance [17], [18]. Studies have shown that YOLOv26 variants can provide excellent accuracy while maintaining lower computational costs compared with previous generations [19], [20]. Despite these advantages, research investigating the use of YOLOv26 Classification for video-based human activity recognition remains very limited.

The UCF101 dataset is one of the most widely used benchmark datasets for evaluating HAR systems. It consists of 13,320 videos distributed across 101 activity categories collected from realistic environments with substantial variations in motion patterns, viewpoints, and backgrounds [5]. Although numerous studies have reported excellent results on UCF101 using CNNs, 3DCNNs, Transformers, and hybrid architectures [21], [8], [9], the performance characteristics of YOLOv26 Classification on this benchmark dataset have not yet been comprehensively analyzed. Furthermore, the influence of frame sampling strategies and model scales on YOLOv26-based activity recognition performance remains unclear.

Therefore, this study aims to analyze the performance of YOLOv26 Classification for video-based human activity recognition using the UCF101 dataset. Six activity classes, namely JumpingJack, Punch, PushUps, Typing, WalkingWithDog, and WritingOnBoard, were selected for evaluation. The proposed framework employs uniform frame sampling to extract representative frames from videos, followed by image classification using YOLOv26 and majority voting to obtain video-level predictions. Comprehensive experiments were conducted using 36 configuration combinations involving four YOLOv26 variants (YOLOv26-n, YOLOv26-s, YOLOv26-m, and YOLOv26-l), three frame sampling settings (8, 16, and 24 frames), and three epoch configurations (50, 100, and 150 epochs). The findings provide a comprehensive evaluation of YOLOv26 Classification for HAR and offer insights into the relationship between model complexity, temporal representation, and recognition performance.

Methodology

This study uses an experimental method to analyze the performance of the YOLOv26 Classification model in recognizing video-based human activities. The dataset used is UCF101, which is one of the benchmark datasets for Human Activity Recognition (HAR) and can be downloaded at https://www.crcv.ucf.edu/data/UCF101.php. Of the 101 activity classes available

in the dataset, this study only uses six activity classes, namely JumpingJack, Punch, PushUps, Typing, WalkingWithDog, and WritingOnBoard. All stages of the study include dataset selection, reading official data sharing files, determining experimental parameters, extracting video frames, forming a classification dataset, training the YOLOv26 model, evaluating at the frame and video level, storing experimental results, and analyzing the results to obtain the best configuration. The overall research flow is shown in Figure 1.

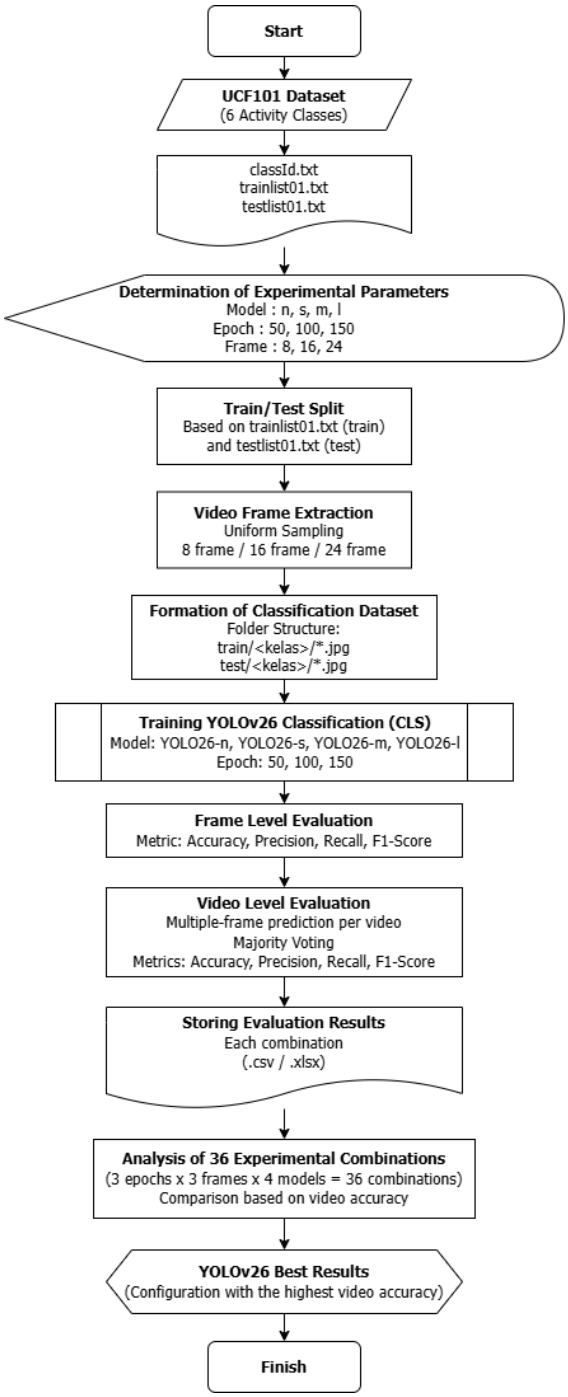

Figure 1. Research methodology flowchart

The number of frames, 8, 16, and 24, was chosen to represent three different levels of temporal resolution. Eight frames were used to represent low temporal information, resulting in lower computational time. Sixteen frames were chosen as an intermediate level, expected to

capture key movement patterns without producing excessive information redundancy. Twenty-four frames were used to test whether increasing temporal information could improve activity classification accuracy. Epoch variations of 50, 100 and 150 were chosen to analyze the effect of training duration on model performance. A value of 50 epochs was used as the initial training baseline, 100 epochs represented intermediate training, and 150 epochs was used to evaluate whether additional learning processes could improve the model's generalization ability or actually cause overfitting.

The model architecture used in this study is based on the YOLOv26 Classification implementation in the Ultralytics framework. Since the official YOLOv26 documentation does not yet provide a detailed visualization of the classification architecture, the model representation is constructed based on the network configuration and component identification results during the training process. Conceptually, the architecture consists of an input image stage, a backbone for feature extraction, a C2PSA (Channel and Spatial Attention)-based feature aggregation module, Global Average Pooling (GAP), a Fully Connected Layer, Softmax, and an output class that generates probabilities for six activity classes. The model architecture representation is shown in Figure 2.

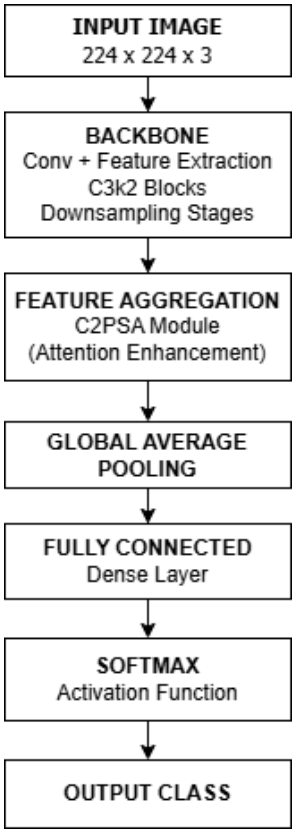

Figure 2. Conceptual architecture of YOLOv26 Classification, the result of interpreting the model implementation in the Ultralytics framework

In addition to the network structure, each YOLOv26 variant has a different model complexity. These differences can be observed from the number of layers, the number of parameters, and the computational requirements expressed in GFLOPs. This information is obtained from the training logs generated during the training process using the Ultralytics framework. Increasing the model size from YOLOv26-n to YOLOv26-l results in an increase in the number of parameters and computational complexity. YOLOv26-n is the lightest model with 1.54 million parameters and a computational requirement of 3.3 GFLOPs, while YOLOv26-l is the largest model with 12.84 million parameters and 49.8 GFLOPs. These differences in model capacity allow for an analysis of the effect of network size on human activity recognition

performance on the UCF101 dataset. A summary of the characteristics of each model is shown in Table 1.

Table 1. Characteristics of the YOLOv26 Classification variant based on the results of model identification in the training process

Model	Layer	Parameter	GFLOPs
YOLOv26-n	86	1.54 M	3.3
YOLOv26-s	86	5.45 M	12.1
YOLOv26-m	106	10.36 M	39.6
YOLOv26-l	176	12.84 M	49.8

All models were trained using the Ultralytics framework with the AdamW optimizer, an initial learning rate of 0.001, a momentum of 0.9, an input image size of 224×224 pixels, and an early stopping mechanism with a patience value of 20 epochs. The batch size was adjusted to the model complexity, namely 8 for YOLOv26-n and YOLOv26-s, 4 for YOLOv26-m, and 2 for YOLOv26-l to avoid GPU memory limitations. Training was run on an NVIDIA GeForce RTX 3050 Laptop GPU with CUDA 12.1.

Results and Discussions

The experimental results code for this research is in here .Based on the evaluation results, there are three configurations that produce the highest video accuracy of 98.26%, namely YOLOv26-m with 16 frames and 50 epochs, YOLOv26-l with 16 frames and 100 epochs, and YOLOv26-l with 16 frames and 150 epochs. Despite having the same accuracy value, the YOLOv26-m configuration shows higher precision, recall, and F1-score values than the two YOLOv26-l configurations. In addition, the model requires fewer epochs, making it more efficient in terms of training time and computational requirements. Therefore, YOLOv26-m with 16 frames and 50 epochs was chosen as the best configuration in this study. The model with the highest accuracy can be seen in Table 2.

Table 2. Results of 36 combinations of YOLOv26 classification experiments

No	Epoch	Frame	Model	Video Accuracy	Video Precision	Video Recall	Video F1-Score
1	50	16	YOLOv26-m	0.983	0.981	0.984	0.982
2	100	16	YOLOv26-l	0.983	0.981	0.982	0.981
3	150	16	YOLOv26-l	0.983	0.981	0.982	0.981
4	100	16	YOLOv26-m	0.978	0.981	0.974	0.976
5	150	8	YOLOv26-s	0.978	0.981	0.974	0.976
6	150	16	YOLOv26-m	0.978	0.981	0.974	0.976
7	50	8	YOLOv26-s	0.970	0.971	0.965	0.966
8	50	16	YOLOv26-s	0.970	0.972	0.965	0.967
9	50	16	YOLOv26-l	0.970	0.967	0.965	0.966
10	50	24	YOLOv26-n	0.970	0.971	0.965	0.966
…	…	…	…	…	…	…	…
36	100	8	YOLOv26-l	0.939	0.951	0.938	0.937

Comparison of video accuracy values obtained from 36 experimental combinations involving variations of the YOLOv26 model, the number of frames, and the number of epochs. In general, all configurations produced high performance with video accuracy values above 93%, indicating that YOLOv26 is able to recognize human activity well on the UCF101 dataset. The best configuration was obtained with the YOLOv26-m model with 16 frames and 50 epochs, which achieved a video accuracy of 98.26%. In addition, it can be seen that the use of 16 frames tends to produce higher accuracy than the use of 8 frames or 24 frames. Although several configurations

of the YOLOv26-l model also achieved similar accuracy values, the YOLOv26-m model was chosen as the best configuration because it was able to provide higher precision, recall, and F1-score values with a fewer number of epochs. These results indicate that the choice of the number of frames and model complexity has a significant influence on the performance of video-based human activity recognition using YOLOv26, as shown in Figure 3.

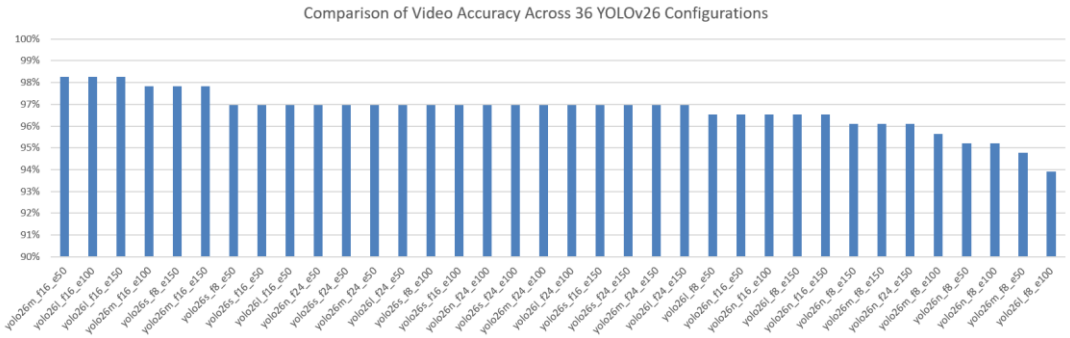

Figure 3. Comparison of video accuracy on 36 experimental combinations consisting of variations of the YOLOv26 model, number of frames, and number of epochs

To better understand the significance of the obtained results, the performance of the proposed YOLOv26 Classification framework was compared with several previous HAR studies conducted on the UCF101 dataset. Jindal et al. [6] reported an accuracy of 96.16% using a feature optimization strategy combined with ResNet101V2, while Ahmad et al. [22] achieved 91.79% using a VGG16-BiGRU architecture. Khare et al. [23] reported an accuracy of 96.23% using a micro-network deep convolutional neural network, and Shin et al. [8] achieved 98.46% using a hybrid ConvNeXt-TCN framework. In this study, the best YOLOv26-m configuration achieved a video accuracy of 98.26%, precision of 98.15%, recall of 98.37%, and F1-score of 98.23%. These results indicate that YOLOv26 Classification is capable of achieving performance comparable to state-of-the-art HAR approaches while employing a simpler pipeline based on frame extraction and majority voting.

The superior performance obtained by YOLOv26-m suggests that model scale plays an important role in balancing representation capacity and generalization ability. Although YOLOv26-l contains a larger number of parameters and higher computational complexity, its performance was not consistently better than YOLOv26-m. In fact, both models achieved the same maximum accuracy of 98.26% under certain configurations, while YOLOv26-m produced slightly higher precision, recall, and F1-score values using fewer training epochs. This finding indicates that increasing model complexity does not necessarily improve recognition performance and may introduce unnecessary computational overhead. Therefore, the medium-scale architecture provides a more favorable trade-off between accuracy and efficiency for the selected HAR classes.

Another important finding is the influence of frame sampling on recognition performance. The results show that configurations using 16 frames generally outperformed those using 8 or 24 frames. This observation suggests that 16 frames provide an optimal balance between spatial and temporal information. When only 8 frames are used, some motion transitions may be omitted, limiting the model's ability to capture complete activity patterns. Conversely, increasing the number of frames to 24 may introduce redundant information between adjacent frames, reducing the contribution of additional temporal information. Similar observations have been reported in previous HAR studies, where excessive temporal redundancy did not always lead to performance improvements [4], [8].

The effectiveness of the proposed majority voting strategy can also be observed from the high video-level performance achieved across all configurations. While classification is performed independently on individual frames, majority voting aggregates frame-level predictions into a more robust video-level decision. This mechanism helps reduce the impact of occasional frame misclassifications caused by motion blur, pose variation, or background complexity. The consistently high video accuracy obtained in this study demonstrates that majority voting is an effective and computationally efficient alternative to more complex temporal modeling approaches such as LSTM, GRU, or Transformer-based architectures. As a result, the proposed

framework offers a practical solution for real-time HAR applications where computational efficiency is an important consideration.

The best model performance analysis was performed using a confusion matrix to evaluate the distribution of predictions for each activity class. The test results showed that most samples were correctly classified, as indicated by the dominance of values on the main diagonal of the matrix. The WritingOnBoard, Typing, WalkingWithDog, Punch, and JumpingJack classes had excellent recognition rates with a relatively small number of errors. The PushUps class still showed some misclassifications, especially against Punch and WalkingWithDog. Nevertheless, overall, the model was able to distinguish the six human activities very well. The confusion matrix results for the best YOLOv26-m model with 16 frames and 50 epochs are shown in Figure 4.

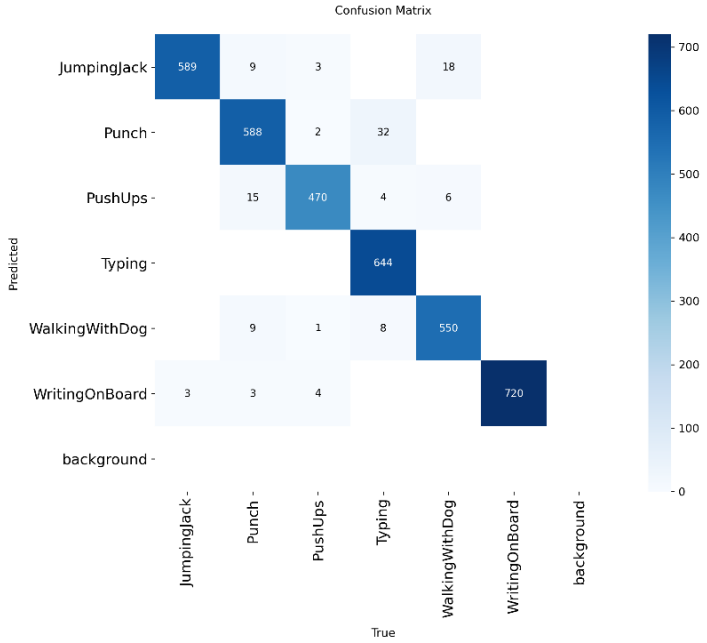

Figure 4. Confusion matrix of human activity classification results using the YOLOv26-m model with 16 frames and 50 epochs

The normalization matrix shows the proportion of correct predictions for each activity class. A value close to 1 indicates that the model has high classification ability for that class. The normalization matrix can be seen in Figure 5.

Figure 5. Normalized confusion matrix of the best YOLOv26-m model with 16 frames and 50 epochs

The confusion matrix analysis further confirms the effectiveness of the proposed framework. Most predictions were concentrated along the main diagonal, indicating that the model successfully distinguished the six activity classes with a low error rate. Activities such as WritingOnBoard and Typing achieved particularly strong recognition performance due to their distinctive visual characteristics and relatively consistent body movements. In contrast, several misclassifications were observed between PushUps and Punch, as both activities involve repetitive upper-body movements that may appear visually similar in certain frames. Similar confusion patterns have been reported in previous HAR studies using UCF101, where activities with overlapping motion characteristics are generally more difficult to distinguish accurately [2], [7]. Nevertheless, the overall number of misclassified samples remained very small, demonstrating the robustness of the YOLOv26 Classification model for video-based activity recognition.

Based on the training curve, the loss function value, indicated by the training loss, experienced a significant decrease from around 0.46 in the initial epoch to nearly 0.02 at the end of training. This decrease indicates that the model is increasingly able to minimize classification errors during the learning process. Meanwhile, the accuracy_top1 and accuracy_top5 values tended to increase and reached a stable condition, indicating that the model has achieved convergence. These results indicate that the model successfully learned human activity patterns without showing any signs of instability during the optimization process. A visualization of the changes in the loss function and accuracy values during training is shown in Figure 6.

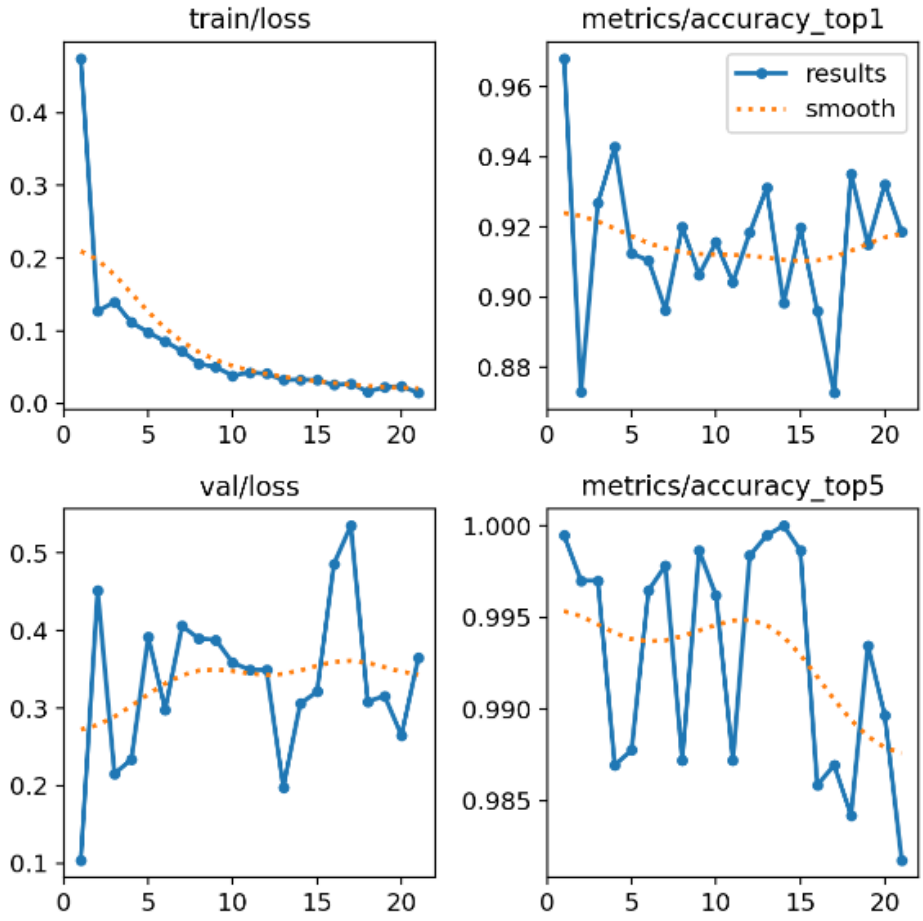

Figure 6. Training and validation curves of the YOLOv26-m model using 16 frames and 50 epochs

To visually illustrate the model's performance, several examples of prediction results on the validation data are shown. The label on each frame indicates the predicted class and the confidence value generated by the model. Each frame shows the predicted activity label and the confidence score generated by the model. The prediction results show that the model is able to recognize activities such as WritingOnBoard, PushUps, and Typing with a high level of confidence,

as indicated by confidence values approaching 1.0 in most samples. Although there are some frames with lower confidence values, such as the PushUps activity, the model is still able to classify the activity correctly. These results indicate that the YOLOv26-m model with 16 frames and 50 epochs has good ability to recognize various human activities on data that was not used during the training process, as shown in Figure 7.

Figure 7. Example of the best prediction results of the YOLOv26-m model with 16 frames and 50 epochs on the validation data

The results showed that using 16 frames resulted in higher accuracy compared to using 8 or 24 frames. This indicates that this number of frames provides an optimal balance between spatial and temporal information. In the 8-frame configuration, some motion information is potentially lost due to too little temporal representation. Conversely, in the 24-frame configuration, there is the possibility of redundant information between frames that does not contribute significantly to the classification process. Therefore, 16 frames is the most effective number in representing the characteristics of human activity in the UCF101 dataset used in this study.

Further research could be conducted by expanding the number of activity classes and using all classes in the UCF101 dataset or larger datasets such as HMDB51 and Kinetics-400 to test the model's generalization ability in more complex scenarios. Furthermore, future research could explore the use of a larger number of frames or more adaptive frame selection methods to better represent temporal information in videos. Developments could also be made by combining YOLOv26 Classification with temporal learning-based models such as Long Short-Term Memory (LSTM), Temporal Convolutional Network (TCN), or video-based Transformer to more optimally utilize inter-frame relationships. Furthermore, testing in real-world environments with varying lighting, shooting angles, backgrounds, and video quality is also necessary to determine the

model's robustness in practical implementation. Finally, further research could evaluate computational efficiency aspects such as inference time, GPU memory usage, and resource consumption on edge computing devices or real-time systems so that the developed model not only has high accuracy but can also be implemented effectively in real operational environments.

Conclusion

This study successfully implemented the YOLOv26 Classification model to recognize six human activities in the UCF101 dataset: JumpingJack, Punch, PushUps, Typing, WalkingWithDog, and WritingOnBoard. The research process involved extracting video frames using a uniform sampling method, creating an image classification dataset, training the YOLOv26 model with four architecture variants (YOLOv26-n, YOLOv26-s, YOLOv26-m, and YOLOv26-l), and evaluating it at the frame and video levels using a majority voting approach. All experiments were conducted on 36 configuration combinations, varying in the number of epochs (50, 100, and 150), the number of frames (8, 16, and 24), and the YOLOv26 model variants.

The test results showed that all configurations were capable of producing high performance with video accuracy values above 93%. Based on an analysis of 36 experimental combinations, the best configuration was obtained by YOLOv26-m with 16 frames and 50 epochs, achieving a video accuracy of 98.26%, a precision of 98.15%, a recall of 98.37%, and an F1-score of 98.23%. These results indicate that the combination of a sufficiently representative number of frames and medium model complexity can produce optimal performance without requiring a larger number of epochs. These findings also indicate that increasing the number of epochs does not always result in a significant increase in accuracy, as some configurations with 100 and 150 epochs produced similar or even lower accuracy values than the best configuration.

A confusion matrix analysis showed that the best model was able to distinguish most activities with a very low error rate. Most predictions were concentrated on the main diagonal of the matrix, indicating that the model successfully recognized the correct class in the majority of the test data. The WritingOnBoard and Typing activities were the easiest to recognize, while relatively more misclassifications occurred for PushUps and Punch, which have similar body movement characteristics across several video frames. However, the number of errors was still very small compared to the number of correct predictions, so it did not affect the overall performance of the model.

The training curve shows that the loss function value consistently decreased during the training process, while the accuracy value tended to increase and reached a state of convergence. This indicates that the model successfully learned the human activity patterns in the dataset without experiencing instability during the optimization process. Furthermore, examples of prediction results on the validation data demonstrate that the model is capable of providing appropriate class predictions with a high degree of confidence across the various human activities tested. Overall, the research results demonstrate that YOLOv26 Classification has excellent capabilities for video-based human activity recognition through a frame extraction and majority voting approach, thus offering potential applications in various applications such as intelligent surveillance systems, human behavior analysis, sports, healthcare, and human-computer interaction.

Acknowledgments

The authors would like to thank Universitas Ahmad Dahlan, Yogyakarta, Indonesia, for the support and facilities provided during the completion of this research.

References

[1] A. Roy, K. D. Gaikwad, A. J. Hude, J. S. Shinde, J. Parekh, and H. R. Nagrani, "Explainable AI (XAI) for object detection and application to satellite imagery," *IEEE Access*, vol. 13, pp. 185597–185623, 2025, doi: 10.1109/ACCESS.2025.3624198.

[2] R. Vrskova, R. Hudec, P. Kamencay, and P. Sykora, "Human activity classification using the 3DCNN architecture," *Appl. Sci.*, vol. 12, no. 2, pp. 1–17, 2022, doi: 10.3390/app12020931.

[3] H. Alshaya, "Neuro-Prismatic video models for causality-aware action recognition in neural rehabilitation systems," *Mathematics*, vol. 14, no. 8, Apr. 2026, doi: 10.3390/math14081341.
[4] M. Li, X. Jiangbo, W. Lu, D. Xinguan, and S. Shuang, "Mamba-YOWO an efficient spatio temporal representation framework for action detection," vol. 48, pp. 1–11, 2026.
[5] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," Dec. 2012, [Online]. Available: http://arxiv.org/abs/1212.0402.
[6] S. Jindal, M. Sachdeva, and A. K. S. Kushwaha, "Quantum behaved intelligent variant of gravitational search algorithm with deep neural networks for human activity recognition Dept . of Computer Science and Engineering I . K . Gujral Punjab Technical University , Jalandhar , India Guru Ghasidas Vishwavi," vol. 50, no. 2, pp. 1–20, 2023.
[7] T. Ahmad, J. Wu, H. S. Alwageed, F. Khan, J. Khan, and Y. Lee, "Human Activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection," *IEEE Access*, vol. 11, no. April, pp. 33148–33159, 2023, doi: 10.1109/ACCESS.2023.3263155.
[8] J. Shin, M. A. M. Hasan, M. Maniruzzaman, S. Nishimura, and S. Alfarhood, "Video-Based human activity recognition using hybrid deep learning model," *C. - Comput. Model. Eng. Sci.*, vol. 143, no. 3, pp. 3615–3638, 2025, doi: 10.32604/cmes.2025.064588.
[9] B. Srinivasaiah and J. Ramegowda, "Human activity recognition using selective kernel network-2D convolutional neural network with ArcFace loss," *IAES Int. J. Artif. Intell.*, vol. 15, no. 1, pp. 350–360, 2026, doi: 10.11591/ijai.v15.i1.pp350-360.
[10] A. A. R. K. Bsoul, "Human activity recognition using graph structures and deep neural networks," *Computers*, vol. 14, no. 1, 2025, doi: 10.3390/computers14010009.
[11] U. M. Butt, H. A. Ullah, S. Letchmunan, I. Tariq, F. H. Hassan, and T. W. Koh, "Leveraging transfer learning for spatio-temporal human activity recognition from video sequences," *Comput. Mater. Contin.*, vol. 74, no. 3, pp. 5017–5033, 2023, doi: 10.32604/cmc.2023.035512.
[12] H. Alshaya, "Neuro-Prismatic video models for causality-aware action recognition in neural rehabilitation systems," *Mathematics*, vol. 14, no. 8, 2026, doi: 10.3390/math14081341.
[13] S. Zahra, N. T. Cao, and T. Kim, "Adaptive text prompting: A training-free framework for zero-shot action recognition," *IEEE Access*, vol. 14, no. February, pp. 45165–45178, 2026, doi: 10.1109/ACCESS.2026.3673165.
[14] Z. V. Isik and T. Ozcan, "ExpAbsTanH: A novel activation function for image and video-based human activity recognition," *IEEE Access*, vol. 14, no. January, pp. 8484–8505, 2026, doi: 10.1109/ACCESS.2026.3653891.
[15] S. Shinde, A. Kothari, and V. Gupta, "YOLO based human action recognition and localization," *Procedia Comput. Sci.*, vol. 133, no. 2018, pp. 831–838, 2018, doi: 10.1016/j.procs.2018.07.112.
[16] M. Elnady and H. E. Abdelmunim, "A novel YOLO LSTM approach for enhanced human action recognition in video sequences," *Sci. Rep.*, vol. 15, no. 1, pp. 1–30, 2025, doi: 10.1038/s41598-025-01898-z.
[17] S. Jing, R. Mai, and X. Gao, "A Method for down quality inspection : YOLO-Based impurity detection and quality quantification," 2026.
[18] C. Julio, F. Silva, C. Del-valle-soto, and C. Bran, "Comparative analysis of previous YOLO detectors and YOLOv26s for real-time weapon detection in video surveillance," no. April, pp. 1–20, 2026, doi: 10.3389/fcomp.2026.1789702.
[19] Y. Aydin, U. Isikdag, S. M. Nigdeli, G. Bekdas, and C. Cakiroglu, "Image-Based Classification of concrete carbonation using YOLO models," *Mater. MDPI*, vol. 19, no. 2198, pp. 1–44, 2026.
[20] H. Keshk and A. Abdallah, "Real-Time UAV-Based oil pipeline and visual anomaly detection using YOLOv26n: A dataset and edge-deployment study," *Drones*, vol. 10, no. 4, 2026, doi: 10.3390/drones10040255.
[21] S. Khan *et al.*, "Enhanced spatial stream of two-stream network using optical flow for human action recognition," *Appl. Sci.*, vol. 13, no. 14, 2023, doi: 10.3390/app13148003.
[22] T. Ahmad, J. Wu, H. S. Alwageed, F. Khan, J. Khan, and Y. Lee, "Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection," *IEEE Access*, vol. 11, pp.

33148–33159, 2023, doi: 10.1109/ACCESS.2023.3263155.
[23] A. Khare, A. Kushwaha, and O. Prakash, "Human activity recognition in a realistic and multiview environment based on two-dimensional convolutional neural network," *J. Artif. Intell. Technol.*, vol. 3, no. 3, pp. 100–107, 2023, doi: 10.37965/jait.2023.0163.