Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065061

ARTICLE

A YOLOv11-Based Deep Learning Framework for Multi-Class Human Action Recognition

Nayeemul Islam Nayeem[1], **Shirin Mahbuba**[1], **Sanjida Islam Disha**[1], **Md Rifat Hossain Buiyan**[1], **Shakila Rahman**[1,*], **M. Abdullah-Al-Wadud**[2] **and Jia Uddin**[3,*]

[1]Department of Computer Science, American International University-Bangladesh, Dhaka, 1229, Bangladesh
[2]Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, 11543, Saudi Arabia
[3]Artificial Intelligence and Big Data Department, Woosong University, Daejeon, 34606, Republic of Korea

*Corresponding Authors: Shakila Rahman. Email: shakila.rahman@aiub.edu; Jia Uddin. Email: jia.uddin@wsu.ac.kr

Received: 02 March 2025; Accepted: 27 June 2025; Published: 29 August 2025

ABSTRACT: Human activity recognition is a significant area of research in artificial intelligence for surveillance, healthcare, sports, and human-computer interaction applications. The article benchmarks the performance of You Only Look Once version 11-based (YOLOv11-based) architecture for multi-class human activity recognition. The article benchmarks the performance of You Only Look Once version 11-based (YOLOv11-based) architecture for multi-class human activity recognition. The dataset consists of 14,186 images across 19 activity classes, from dynamic activities such as running and swimming to static activities such as sitting and sleeping. Preprocessing included resizing all images to 512 × 512 pixels, annotating them in YOLO's bounding box format, and applying data augmentation methods such as flipping, rotation, and cropping to enhance model generalization. The proposed model was trained for 100 epochs with adaptive learning rate methods and hyperparameter optimization for performance improvement, with a mAP@0.5 of 74.93% and a mAP@0.5-0.95 of 64.11%, outperforming previous versions of YOLO (v10, v9, and v8) and general-purpose architectures like ResNet50 and EfficientNet. It exhibited improved precision and recall for all activity classes with high precision values of 0.76 for running, 0.79 for swimming, 0.80 for sitting, and 0.81 for sleeping, and was tested for real-time deployment with an inference time of 8.9 ms per image, being computationally light. Proposed YOLOv11's improvements are attributed to architectural advancements like a more complex feature extraction process, better attention modules, and an anchor-free detection mechanism. While YOLOv10 was extremely stable in static activity recognition, YOLOv9 performed well in dynamic environments but suffered from overfitting, and YOLOv8, while being a decent baseline, failed to differentiate between overlapping static activities. The experimental results determine proposed YOLOv11 to be the most appropriate model, providing an ideal balance between accuracy, computational efficiency, and robustness for real-world deployment. Nevertheless, there exist certain issues to be addressed, particularly in discriminating against visually similar activities and the use of publicly available datasets. Future research will entail the inclusion of 3D data and multimodal sensor inputs, such as depth and motion information, for enhancing recognition accuracy and generalizability to challenging real-world environments.

KEYWORDS: Human activity recognition; YOLOv11; deep learning; real-time detection; anchor-free detection; attention mechanisms; object detection; image classification; multi-class recognition; surveillance applications

1 Introduction

Human activity recognition (HAR) has been a primary challenge of artificial intelligence, with applications of a crucial nature in surveillance, healthcare, sports, and human-computer interaction. This research aims to compare four YOLO-based models—YOLOv8, YOLOv9, YOLOv10, and the proposed YOLOv11—for human activity classification of 19 different human activities, from dynamic activities (e.g., running, swimming) to static ones (e.g., sitting, sleeping). The models are trained and tested on a handpicked Kaggle dataset with 14,186 labeled images. Accurate HAR in natural environments has several challenges: Subtle Patterns of Activity: People's behavior typically involves subtle patterns and combined motions, making it difficult to distinguish between similar actions, such as sleeping and sitting, or drinking and eating. Real-time Processing Requirements: Some applications, e.g., surveillance and interactive ones, necessitate real-time processing without a loss of precision. Computational Efficiency: The accuracy of detection must be traded off against computation cost, particularly when edge-device deployments are involved. Dynamic Environmental Conditions: Real environments place variations in illumination, occlusions, and concurrent activities on the recognized objects that complicate recognition. In an attempt to alleviate such conditions, we employed and compared progressive YOLO structures with a focus on the improved functionality of YOLOv11 in handling complex activity recognition tasks. Architectural advancements in YOLOv11 include improved feature extraction and attention mechanisms to boost performance across all metrics. Models were trained for 50 and 100 epochs on input images of fixed sizes (512 × 512 pixels) with consistent hyperparameters to enable an equivalent test. YOLOv11 performed better than all its earlier iterations on all the test measures, including precision, recall, and mean Average Precision (mAP), and thus is fit for real-world application. This work opens the door to future research in solid, multi-class HAR systems. Subsequent research will aim to narrow the gap between theory and application, enhance model generalizability and performance in dynamic and diverse environments. The paper [1] proposes the use of YOLOv3 with transfer learning for real-time object detection and tracking in surveillance applications. It is trained on the MS COCO dataset and has an accuracy of 99.71% and mAP of 61.5. The system is tuned to handle dynamic scenarios effectively. Even though it works well, it lacks accuracy and contextual awareness. The authors recommend exploring newer versions of YOLO and improved contextual modeling for future studies. This paper [2] proposes a deep convolutional neural network (DCNN) for human activity recognition using micro-Doppler radar with an accuracy rate of 97.5%. Compared to traditional approaches like SVM and Decision Trees, it is more resistant to environmental noise and varying lighting conditions. The model is able to distinguish between walking, running, and jumping successfully. Areas of future work include enlarging the dataset, including real-time solutions, and developing a hybrid model that combines radar with visual or thermal data. The paper demonstrates the feasibility of DCNN and micro-Doppler radar for human action recognition in moving scenes. Human action recognition [3] from static images proves challenging due to variations in postures and backgrounds. Advances through Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), and Faster R-CNN have improved the performance of the model. The models use local and global predictions, achieving higher accuracy on benchmarks like Stanford40 and PASCAL VOC. Multimodal data fusion and making the model more efficient are areas of research that can be focused on in the future. This work [4] presents a human action recognition system utilizing YOLO for human detection and OpenPose for pose estimation. YOLO detects humans from videos, and OpenPose estimates skeletal frames to classify actions with improved performance under complex backgrounds and occlusions. The technique improves action classification in real time. The paper [5] introduces YOLO-face, a face detection model based on YOLOv3 that addresses the issue of face scale variation. It uses the WIDER FACE dataset, anchor box optimization, a purer loss function, and a Feature Pyramid Network for improved accuracy. YOLO-face achieves high Average Precision (AP) and a 38 FPS processing rate, rendering it real-time capable. Research work can be steered in the direction of architectural enhancement and broader application for increased efficiency and precision. This paper [6] surveys Human Action Recognition (HAR) methods in RGB and skeleton modalities, listing their strengths and limitations. RGB methods offer rich spatial information but are affected by lighting and noise, while skeleton methods contain robust motion information without supplying context. It surveys datasets like UCF101, Kinetics, NTU RGB+D, and MSR Action3D, listing issues like dataset variability. The work [7] proposes a multi-scale convolution transformer model for human activity recognition, combining CNNs and transformers to learn local and global temporal features. It uses the Channel State Information (CSI) dataset to identify and timestamp activities with high micro F1-scores of 98.37% for weak activities and 92.81% for strong activities. The hybrid model effectively handles complicated activity sequences. The paper [8] proposes a multimodal human activity recognition (HAR) system using skeleton and RGB data to achieve state-of-the-art performance on challenging datasets. It fuses skeleton features and RGB motion dynamics using decision-level fusion with classifiers like Random Forest and SVM, achieving 98.8% precision on CAD-60. This survey [9] elaborates on the impact of deep learning on human activity recognition (HAR) and its applications in health care, sports, and smart spaces. It elaborates on datasets, techniques like CNNs, RNNs, and hybrid approaches, and performance metrics, noting better performance of CNNs with spatial data and RNNs with temporal data. This work [10] presents a human activity recognition (HAR) method with low-cost Doppler radar sensors and Convolutional Neural Networks (CNNs). Radar sensors detect micro-Doppler signatures, which are thereafter converted to spectrograms to extract features for activity recognition such as walking, sitting, and standing. The system achieves high classification accuracy using a low-cost, non-intrusive solution. The approach is promising for applications in healthcare, security, and smart homes. The key contributions of this paper, concerning human activity recognition, are as follows:

1. In this work, an enhanced YOLOv11 architecture for real-time human activity detection and object recognition across diverse scenarios has been proposed.
2. First, the input data, consisting of images and videos, is preprocessed using techniques like resizing, normalization, rotation, and flipping to ensure robustness under varying conditions.
3. Second, the proposed model is trained on a curated dataset of 14,186 images representing 19 distinct activity classes, utilizing advanced training techniques such as dynamic label assignment and self-distillation to enhance generalization. The proposed YOLOv11 model incorporates advanced features such as C3k2 Blocks, SPPF, and C2PSA attention mechanisms for improved feature extraction and activity recognition.
4. Finally, the performance of YOLOv11 is evaluated and compared with baseline models, such as YOLOv8, YOLOv9, and YOLOv10, as well as ResNet50 and EfficientNet. The results demonstrate that YOLOv11 outperforms all baseline models in accuracy, precision, recall, and real-time processing speed, making it suitable for applications like surveillance, healthcare, and sports analysis.

2 Literature Review

Human activity recognition is increasingly important for surveillance, healthcare, and human-computer interaction applications. The paper reviews the recent HAR methods, classifying them as YOLO-based and non-YOLO methods, and compares them with the proposed YOLOv11 architecture. This paper [11] proposes a radar-based HAR model using micro-Doppler features, DCNNs, LSTMs, and attention mechanisms with over 85% mAP performance. However, it does not support real-time processing, multimodal fusion, explainability, and edge deployment optimization, limiting its application in real-time scenarios. The paper [12] introduces a skeleton-based hybrid HAR model combining CNNs to learn spatial features and LSTMs to capture temporal patterns, enhancing precision, accuracy, and F1-score. In contrast to conventional techniques, which deal with these features separately and suffer from complicated tasks, the novel model learns these features simultaneously, demonstrating enhanced generalizability on the UTD-MHAD dataset. This paper [13] explores RGB-D sensing for HAR by combining RGB, depth, and skeleton data for more powerful spatial-temporal analysis using models like Two-Stream Networks, C3D, and ST-GCN. The paper [14] presents a two-stream convolution-augmented transformer network for improving WiFi-based CSI data processing for HAR to address the weak performance of the traditional RNN and LSTM models with inferior processing of long-term the rich spatial-temporal information in CSI data. The paper [15] introduces a dynamic representation and sequence matching-based HAR approach with RGB-D image skeleton features, addressing the issue that most existing HAR approaches cannot effectively capture temporal structures and pose variations, particularly for activities with variable speed or subtle motion distinctions. By utilizing shape dynamic time warping (shapeDTW), it performs more accurately than existing methods on three public datasets and enhances robustness in detecting such activities. The paper [16] presents an HAR system using angle inclination-based HAR approach and keypoints descriptor network to represent temporal relationships between key poses and address the drawbacks of many current state-of-the-art systems that are restricted in the representation of pose transitions and motion variations, especially when there are varying speeds of activities. The paper [17] presents a HAR system based on CNN using pose-based keypoint features to provide higher accuracy and counter the disadvantage of current models in not dealing with detailed pose features in clutter or occlusion scenarios. It has OpenPose and COCO for pose estimation but lacks temporal modeling, implemented in future work using LSTMs or transformers. The paper [18] presents a real-time HAR system with YOLOv5 and Tiny YOLO, which is trained using COCO and VOC datasets, to detect undesired activities like violence or robbery with improved accuracy in heterogeneous scenarios. The limitation of current systems is that they are not capable of working in a real-time context-aware activity recognition due to variation in environmental conditions. The research [19] proposes the novel YOLOv8 to detect human activity using CSP networks, data augmentation, and optimum IoU thresholds, with experimentation on datasets like COCO, AVA, and Kinetics. The paper [20] proposes a hybrid action recognition framework combining 3DCNN, Spatial Depth-Based Non-Local layers, and Deep Capsule Networks to learn global temporal and local spatial information. Current methods do not model long-range spatiotemporal dependencies, limiting performance on complex tasks, which the proposed model addresses. The paper [21] introduces a HAR system using physiological sensors and CNNs and LSTMs for spatial-temporal pattern learning from datasets like RealWorld HAR, PAMAP2, and UCI HAR. It addresses sensor quality and activity complexity limitations through attention mechanisms, multimodal fusion, and optimized architecture to achieve state-of-the-art performance. The paper [22] presents a real-time obstacle detection method using YOLOv8 on UAV aerial images with a 96% F1 score at 200 epochs. It addresses the shortcomings of existing methods, where there is a lack of real-time processing and precise detection of obstacles, particularly in complex environments, improve navigation reliability for safety-critical applications. The paper [23] traces the evolution of YOLO models for object detection in UAVs with increasing speed, accuracy, and real-time performance. Computational efficiency and accuracy in difficult situations are challenges, detection time without optimizing performance in difficult situations. The survey [24] explains YOLO models like YOLOv8 and 3D-YOLO for human action recognition considering accuracy, temporal comprehension, and parameters of significance. Despite the advancements, most YOLO models are spatial and not temporal, and therefore are issues in real-time processing, multimodal fusion, and edge deployment, with instances of datasets like UCF101, Kinetics-600, and AVA being put forward. The paper [25] presents a ConvNet-based HAR model that integrates dynamic and RGB images with 98.5% accuracy on the KTH dataset. Traditional HAR methods based on static or depth images are plagued by the lack of temporal context, generalizability to complex environments, and limited resources on edge devices, which are addressed in the proposed model. The paper [26] proposes an improved YOLOv5 model for detecting small objects, with improved architecture, anchor box optimization, and attention mechanisms, achieving good performance on DOTA, VisDrone, and COCO benchmarks. However, challenges like low-resolution images, occlusion, and complex backgrounds remain. The paper [27] explores ensemble learning for human activity recognition on pre-trained CNNs like ResNet, VGG, and EfficientNet on UCF101, Kinetics-700, and NTU RGB+D datasets. However, it remains difficult to strike a balance between the capacity of a model and computational cost, particularly in real-time applications, with future research focusing on scalability and data dependency. The paper [28] presents YOLOv3 for human activity recognition from high-resolution aerial images with 85.3% mAP@0.5 and 25 FPS. The paper [29] traces YOLO from YOLOv1 to YOLOv5 and observes the advancement in speed, accuracy, and efficiency with the help of anchor boxes, learning rates, and datasets like COCO and PASCAL VOC. All development aside, object continuity and real-time tracking in low-resource dynamic environments pose challenges that require optimization.

The proposed YOLOv11 model significantly differs from other prior YOLO and non-YOLO methods in the integration of advanced architectural elements like C3k2 Blocks, SPPF, and C2PSA. The model improves over prior limitations in detecting overlapping or harder activities and achieves a high mAP@0.5 score of 74.93%. The model is trained on a curated, augmented dataset with 14,186 images for 19 diverse activity classes. Unlike skeleton or radar-based systems specialized in a fixed, narrow task, YOLOv11 performs well in both static and dynamic scenes. It is optimized for real-time with an 8.9 ms inference time per image. Its light architecture means it is accurate while maintaining deployability on resource-limited edge devices.

3 Methodology

3.1 Image Acquisition

The dataset contains 14,186 images across 19 activity classes, including dynamic (e.g., running, swimming) and static (e.g., sitting, sleeping) activities. These were taken from Kaggle; further, several other frames were generated by using videos for further different views and contexts. Class imbalance was handled by augmenting the data of the underrepresented activities like boxing and cooking, through extracting video frames. The images were all annotated in YOLO's bounding box format for consistency and compatibility with the training pipeline. Low-quality or irrelevant images were removed to maintain dataset quality. This curated dataset provided a solid foundation for training YOLOv11 and related models. Fig. 1 shows some sample image datasets.

3.2 Image Pre-Processing

The dataset of 14,186 images required data pre-processing for effective training and performance evaluation of the proposed YOLOv11 and other models. All the images were resized to 512 × 512 pixels for input consistency. The data was split 70/15/15 for training, validation, and testing with an equal proportion of all 19 activity classes. Low-quality or irrelevant images were excluded. Various augmentation techniques were applied to simulate real-world variations and enhance model generalization, including random flipping, rotation, cropping, and brightness adjustment. Two specific augmentation cases were used, as detailed in Case 1: The images were randomly lightened ±25% to simulate changing light and facilitate visibility. This step ensures that the dataset includes a range of lighting conditions, thereby enhancing model robustness. Case 2: Applied random flipping—both horizontal and vertical. This helped the model learn better from mirrored orientations likely to occur during inference.

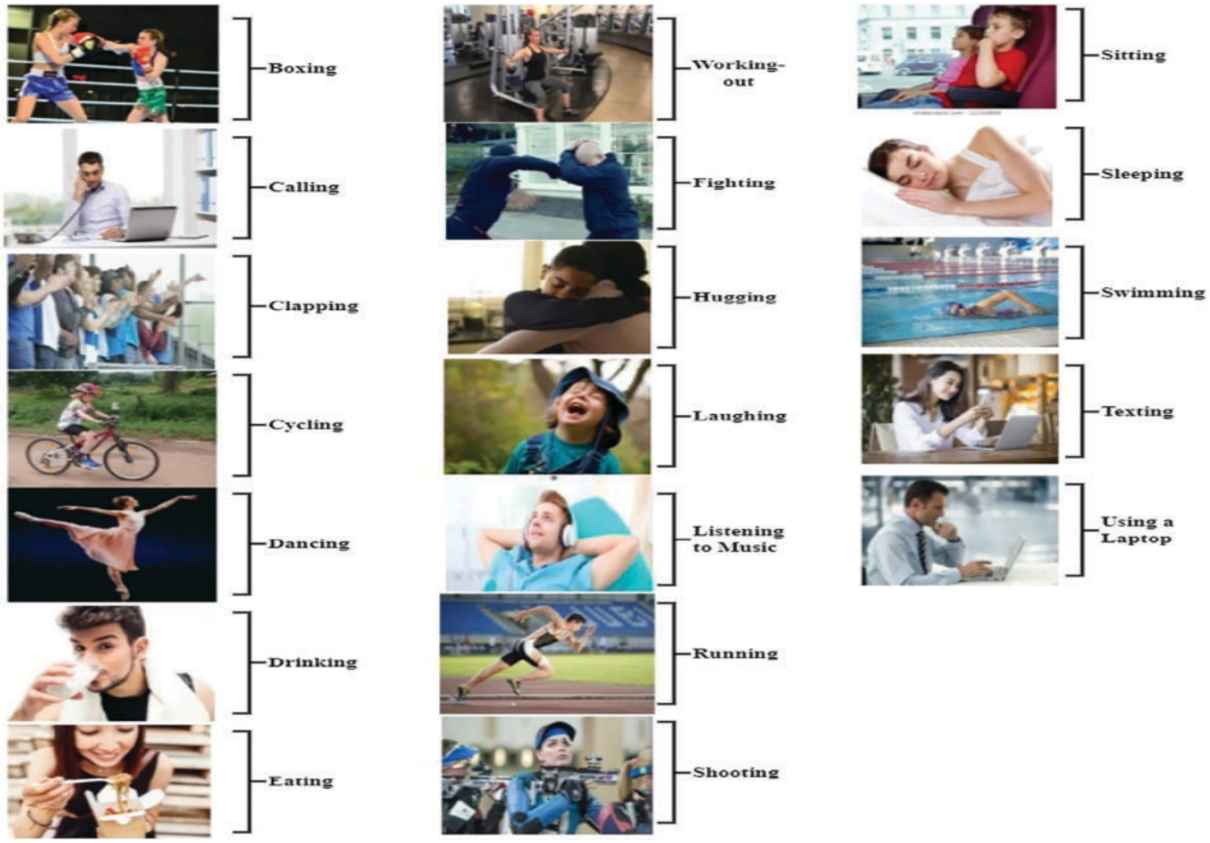


Figure 1: Sample image datasets from the curated human activity recognition dataset

Table 1: Dataset statistics

Types of activity	Quantity	Types of activity	Quantity
Boxing	370	Laughing	681
Calling	617	Listening to music	723
Clapping	792	Running	817
Cycling	548	Shooting	407
Dancing	795	Sitting	836
Drinking	782	Sleeping	716
Eating	713	Swimming	579
Working-out	686	Texting	479
Fighting	677	Using a laptop	635
Hugging	826	**Total**	**14,186**

Table 1 shows the distribution of image samples across the 19 activity classes in the dataset. YOLOv11 utilized Mosaic Augmentation, which combines multiple images to simulate overlapping activities and improve robustness. Augmentations enriched the dataset with more real-world-like, diverse, and high-quality data. These steps significantly improved YOLOv11's performance in identifying human activities in dynamic and static environments. The training, validation, and evaluation benefited from this diverse dataset, resulting in more accurate, less biased, and better generalized models. Fig. 2 presents some examples of applied augmentations and their contributions to dataset enrichment.

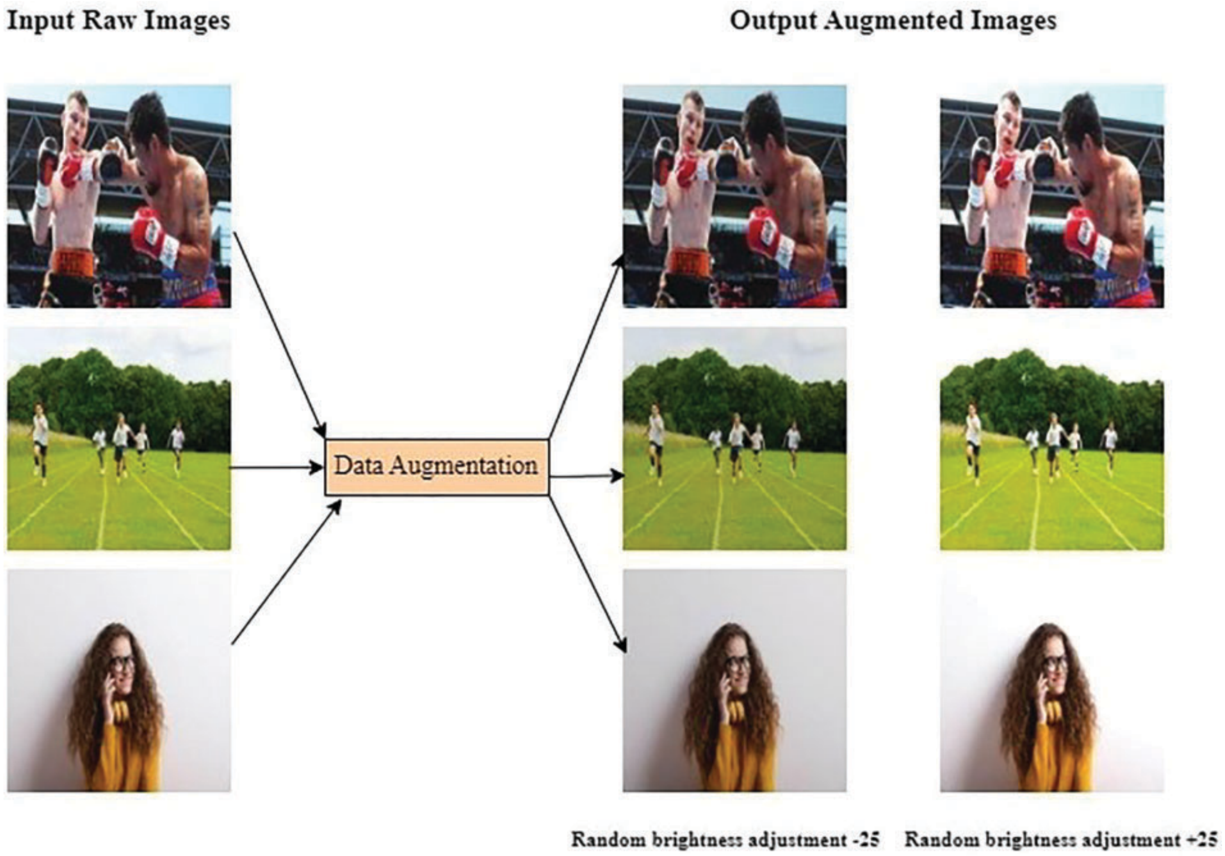

Input Raw Images
Output Augmented Images
Data Augmentation
Random brightness adjustment -25
Random brightness adjustment +25

Figure 2: Examples of applied augmentations enriching the dataset

3.3 Image Resizing and Labeling

YOLO format rescaling and annotation were important to enable training YOLOv11 and other models. All 14,186 images were resized to 512 × 512 pixels for a balance between computational cost and maintaining detail in order to identify activities correctly. Images were annotated using bounding boxes and labeled as one of 19 categories of human activity, such as dynamic (running, swimming) and static (sitting, sleeping) actions, with Roboflow. Fig. 3 illustrates this process: raw images (left), annotated samples with bounding boxes (center), and visualization layers for segmentation and classification (right), facilitating model interpretability. Proper, visual-content-based annotation facilitated effective object localization and classification. These labels were the foundation of supervised learning, allowing YOLO models, whether anchor-free or anchor-based, to learn spatial and contextual activity features, which facilitated robust multi-class human activity recognition within the dataset.

3.4 Model Architecture

The YOLOv11 structure surpasses earlier YOLO frameworks in numerous aspects, enabling more effective real-time human activity recognition across five optimized stages. Phase 1 involves image normalization and Mosaic-based augmentation, enhancing preprocessing for better compensation of overlapping and heterogeneous activity conditions. Phase 2 introduces new feature blocks composed of 3 × 3, 5 × 5, and dilated convolutions that excel over the typical C3 blocks in YOLOv8-v10 in capturing fine-grained and wide-range features. Phase 3 integrates SPPF and C2PSA modules for multi-scale feature fusion and attention-aware, improving detection of subtle motion in complex scenes. Phase 4 refines the prediction head by using C3k2 and CBS layers in P3–P5 scales, generating more accurate bounding boxes and stable outputs in dense or dynamic environments. Phase 5 strengthens post-processing with more effective non-maximum suppression and context-aware adjustments to reduce false positives. These enhancements collectively yield greater mAP, faster inference speed, and more robustness compared to earlier YOLO variants.

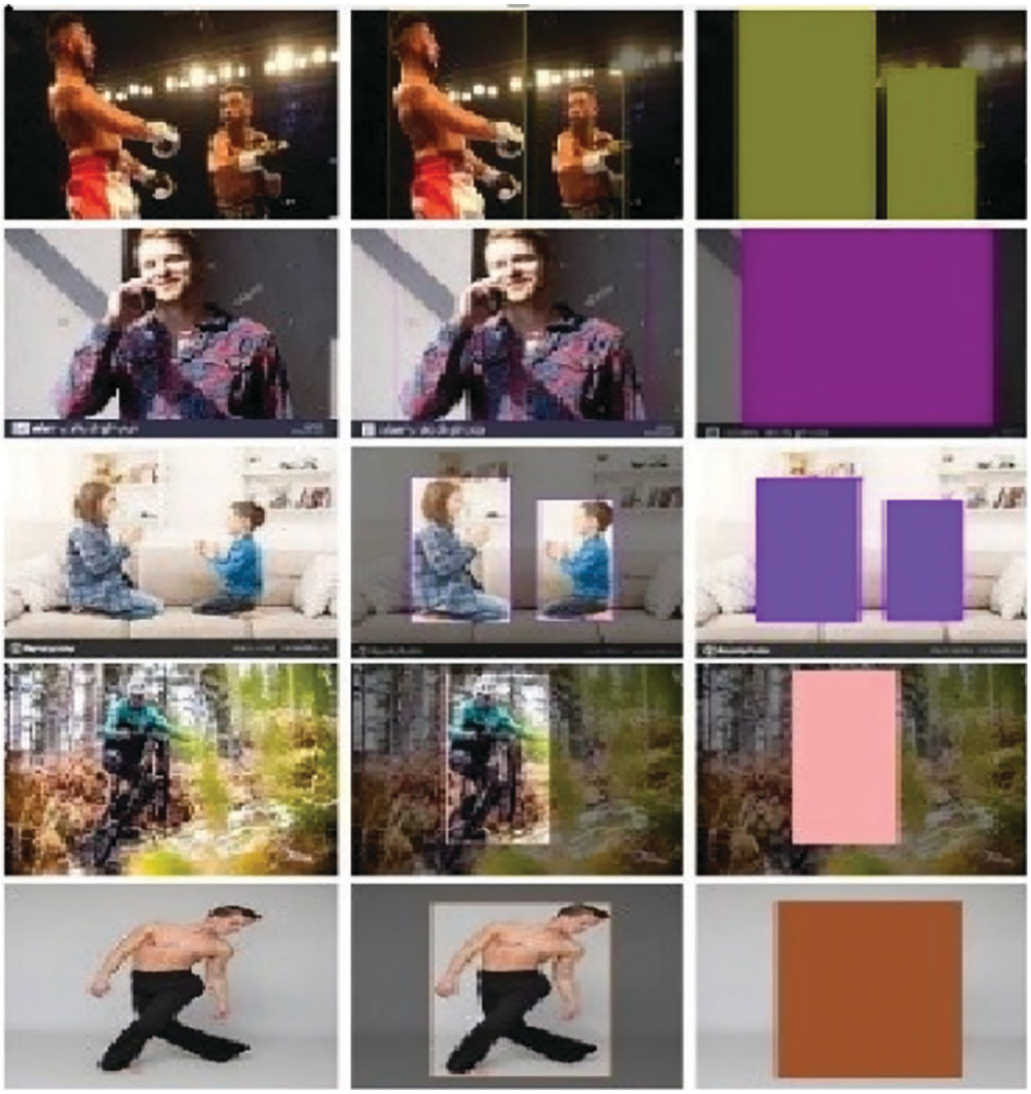


Figure 3: Sample labeled data

In this study, YOLOv8, YOLOv9, YOLOv10, YOLOv11, ResNet50, and EfficientNet were trained to identify human activity in nineteen various activity classes, including boxing, dancing, sleeping, and using a laptop. A null class was also considered for images without relevant context but not included in principal class consideration. The training dataset consisted of approximately 14,186 labeled images collected from diverse sources such as Google, GitHub, and Roboflow, depicting a broad range of real-world activity settings. All models were trained with a consistent learning rate of 0.01 and weight decay of 0.001, a batch size of 16, and an image size of 512 × 512, set constant for all to ensure consistency in evaluation. YOLO and ResNet50 models were trained for 50 to 100 epochs, while EfficientNet, due to its lightweight nature and considering the computational constraints, was trained for 50 epochs. It was trained to predict both bounding boxes and class labels of input images accurately, with weight updating based on the labeled dataset. Although these settings provided good baseline performance, certain limitations still existed. Due to the hardware constraints, the training was restricted to 100 epochs, and advanced augmentation techniques like Mosaic and MixUp were not used. As future work, it is suggested to try more training epochs, use advanced augmentation techniques, and utilize adaptive learning rate schedules like cosine annealing to further improve generalization and performance. Fig. 4 shows proposed YOLOv11 architecture.

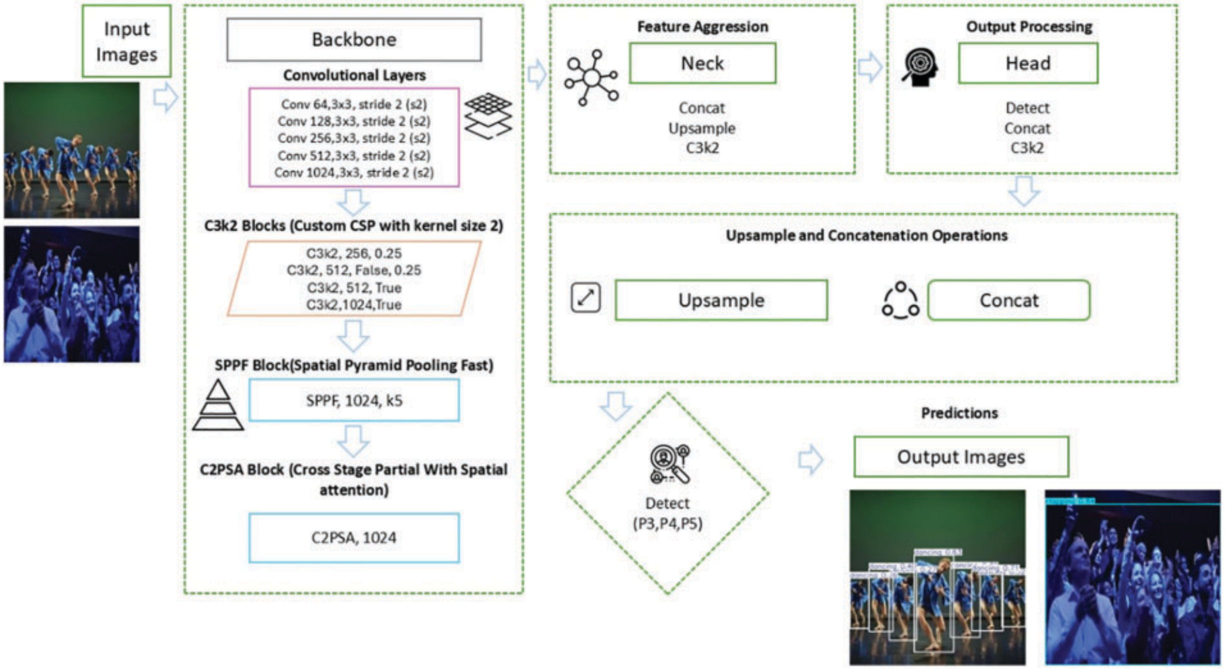

Input Images
Backbone
Convolutional Layers
Conv 64,3x3, stride 2 (s2)
Conv 128,3x3, stride 2 (s2)
Conv 256,3x3, stride 2 (s2)
Conv 512,3x3, stride 2 (s2)
Conv 1024,3x3, stride 2 (s2)
C3k2 Blocks (Custom CSP with kernel size 2)
C3k2, 256, 0.25
C3k2, 512, False, 0.25
C3k2, 512, True
C3k2,1024,True
SPPF Block(Spatial Pyramid Pooling Fast)
SPPF, 1024, k5
C2PSA Block (Cross Stage Partial With Spatial attention)
C2PSA, 1024
Feature Aggression
Neck
Concat
Upsample
C3k2
Output Processing
Head
Detect
Concat
C3k2
Upsample and Concatenation Operations
Upsample
Concat
Detect
(P3,P4,P5)
Predictions
Output Images

Figure 4: Proposed YOLOv11 architecture

These training configurations thus gave a ground on which the models learn effectively to perform suitably on a wide variety of human activities. All models shared the same parameters and preprocessing; hence, any difference in performance between these models can be due to their architecture and learning only. The complete set of training parameters used for YOLOv11 is listed in Table 2.

Table 2: YOLOv11 model parameter

Parameter	Value
Batch size	16
Number of epochs	100
Optimizer	Adamw

(Continued)

Table 2 (continued)

Parameter	Value
Pre-trained	COCO model
Learning rate	0.01
Weight decay	0.001
Patience	50

These training configurations thus gave a ground on which the models learned effectively to perform suitably on a wide variety of human activities. All models shared the same parameters and preprocessing; hence, any difference in performance between these models can be due to their architecture and learning only. The complete set of training parameters used for YOLOv11 is listed in Table 2.

3.5 Model Evaluation

The evaluation of the YOLOv11 model is observed to be strong in detecting and classifying 19 human activity classes, both dynamic, like running and swimming, and static, like sitting and using a laptop. Training and testing were performed in a Google Colab environment on a Tesla T4 GPU with 16 GB of memory. As reported in Table 3, the model is composed of 512 layers, has 35,812,450 parameters, and a computational complexity of 120.7 GFLOPs. It achieved an mAP@0.5 of 74.49% and an mAP@0.5-0.95 of 63.17%, higher than previous YOLO models. Class-wise evaluation noted high precision and recall across all activities, with superior performance in clapping (Precision 0.81, Recall 0.79), drinking (Precision 0.75, Recall 0.77), and use of a laptop (Precision 0.85, Recall 0.83). The model generalizes well in both high-motion and static classes and performs well in diverse situations. The inference pipeline is also quick, with only 0.18 ms for pre-processing, 8.9 ms for model inference, and 1.8 ms for post-processing per image. The results confirm that YOLOv11 offers improved multi-class activity recognition without compromising on the speed and efficiency needed for real-time application in real-world scenarios.

Table 3: YOLOv11 model parameter

Parameter	Value
Model layers	512
Model parameters	35,812,450
Gradients	35,812,429
GFLOPs	120.7

4 Result Analysis

The performance analysis of YOLO models (YOLOv8, YOLOv9, YOLOv10, and YOLOv11), ResNet50, and EfficientNet across 50 and 100 epochs provides insights into their strengths and limitations in detecting and classifying complex human activities. The evaluation metrics, including F1 score, mAP@0.5, and mAP@0.5-0.95, highlight their learning, generalization, and detection capabilities. The comparative testing performance of YOLOv11, YOLOv8, YOLOv9, YOLOv10, ResNet50, and EfficientNet across various metrics and training epochs is presented in Table 4.

Table 4: Testing performance of YOLOv11, v10, v9, v8, ResNet50 and EfficientNet

Model	Epoch	Class	Trainable parameters	F1 score	mAP@0.5	mAP@0.5-0.95
Proposed YOLOv11	50	All (19)	~35.8M	0.71	0.74492	0.63173
Proposed YOLOv11	100	All (19)	~35.8M	0.71	0.74939	0.64108
YOLOv10	50	All (19)	~23.4M	0.69	0.72261	0.60712
YOLOv10	100	All (19)	~23.4M	0.69	0.72506	0.61014
YOLOv9	50	All (19)	~12.5M	0.70	0.72928	0.60947
YOLOv9	100	All (19)	~12.5M	0.71	0.73649	0.62638
YOLOv8	50	All (19)	~4.7M	0.6971	0.72337	0.6019
YOLOv8	100	All (19)	~4.7M	0.70	0.7317	0.60557
ResNet50	50	All (19)	~25.6M	0.25	0.23	0.18
ResNet50	100	All (19)	~25.6M	0.26	0.25	0.2
EfficientNet	50	All (19)	~5.3M	0.15	0.1	0.08
EfficientNet	100	All (19)	~5.3M	0.17	0.13	0.09

Fig. 5 shows that YOLOv11 outperformed all other models on all evaluative measures. On 50 epochs, it performed an F1 score of 0.71, a mean Average Precision (mAP) at 0.5 of 74.49%, and an mAP ranging from 0.5 to 0.95 of 63.17%. These measures showed a minimal improvement by 100 epochs. The model showed strong precision and recall for all 19 classes of activities. For dynamic activities like boxing and dance, precision measures were above 85%. In addition, consistent validation losses within initial training resulted in minimal overfitting and misclassifications, supported by a confusion matrix (Fig. 6).

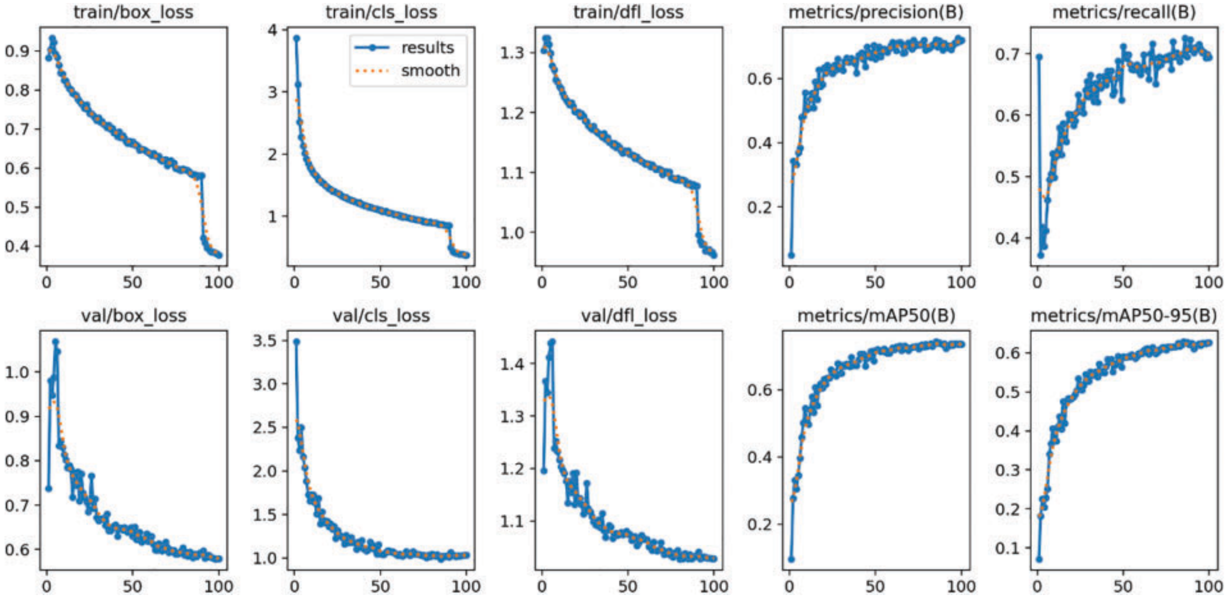
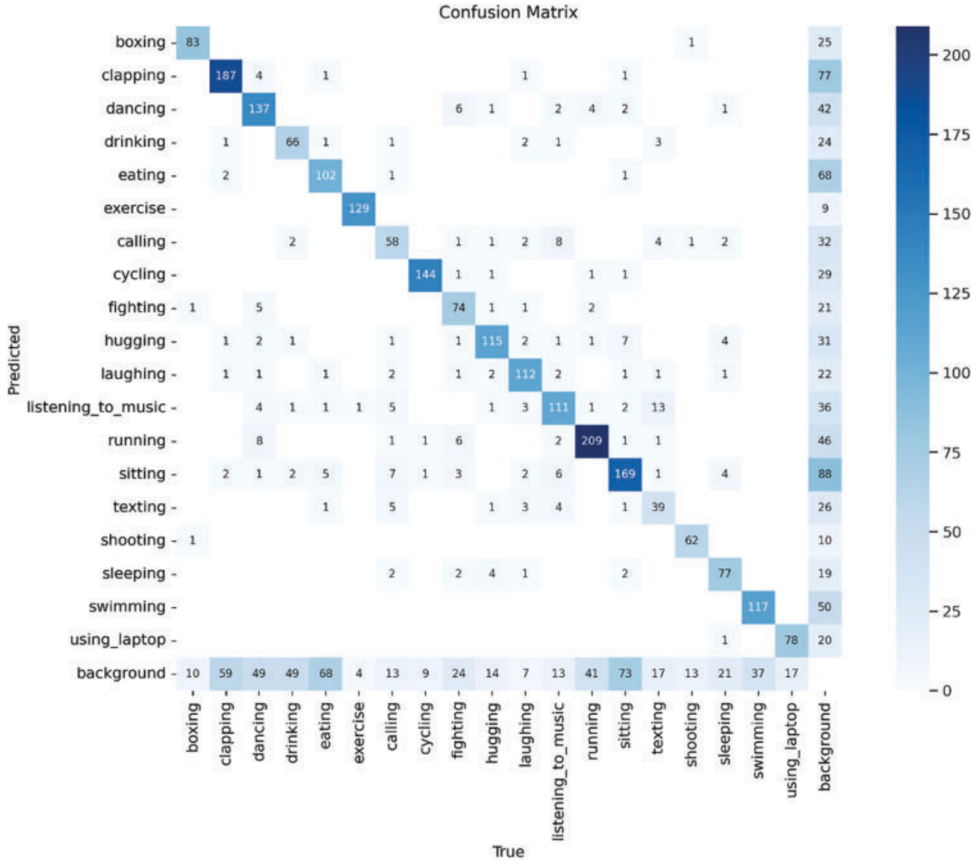


Figure 5: Training curve based on YOLOv11 for 100 epochs



	boxing	clapping	dancing	drinking	eating	exercise	calling	cycling	fighting	hugging	laughing	listening_to_music	running	sitting	texting	shooting	sleeping	swimming	using_laptop	background
boxing	83															1				25
clapping		187	4		1						1			1						77
dancing			137						6	1		2	4	2			1			42
drinking		1		66	1		1				2	1			3					24
eating		2			102		1							1						68
exercise						129														9
calling				2			58		1	1	2	8			4	1	2			32
cycling								144	1	1			1	1						29
fighting	1		5						74	1	1		2							21
hugging		1	2	1			1		1	115	2	1	1	7			4			31
laughing		1	1		1		2		1	2	112	2		1	1		1			22
listening_to_music			4	1	1	1	5			1	3	111	1	2	13					36
running			8				1	1	6			2	209	1	1					46
sitting		2	1	2	5		7	1	3		2	6		169	1		4			88
texting					1		5			1	3	4		1	39					26
shooting	1															62				10
sleeping							2		2	4	1			2			77			19
swimming																		117		50
using_laptop																	1		78	20
background	10	59	49	49	68	4	13	9	24	14	7	13	41	73	17	13	21	37	17	

Figure 6: Confusion matrix of YOLOv11-based model for 100 epochs

The improvements observed in the performance metrics of YOLOv11 are inherently attributed to its architecture adjustments. Unlike traditional backbone architectures, YOLOv11 features C3k2 modules, which combine multiple convolutional paths through the use of dilated filters. This adjustment broadens the receptive field, enabling more complex spatial feature extraction, which is essential for recognizing overlapped and context-dependent human movements. In addition, YOLOv11 features the inclusion of the C2PSA (Cross-Stage Partial Self-Attention) mechanism, which enhances its focus on body regions relevant for actions while ignoring irrelevant details of the background. This ability allows the model to clearly distinguish between similar postures, including sitting and sleeping. The SPPF module is a landmark improvement, allowing for effective multi-scale contextual data integration. When complemented by a sophisticated neck architecture, which combines lower-level and semantic features, YOLOv11 is shown to perform high-quality detection even for densely populated scenarios with a high number of individuals. This provides a better performance balance between dynamic and static categories. Consequently, YOLOv8 does not possess elaborate attention mechanisms and struggles with handling overlapped stationary actions. The speed-optimized YOLO-NAS variant is particularly focused on detecting object instances with smaller sizes, yet it is insufficient for supplying ample spatial depth for deep analysis of entire body movements. Although YOLOv10 and YOLOv9 are improvements over previous ones, neither meets the level of generalization and inference performance attained by YOLOv11. YOLOv11's superior performance can be largely attributed to improvements in its architectural design that include C3k2 blocks, SPPF modules, and the C2PSA attention mechanism. Together, these components have improved spatial feature extraction, attentional mechanisms, and multi-scale representation, contributing to the model's improved ability to detect complex human actions.

Unlike its predecessors, like YOLOv8, without attention refinement, and YOLO-NAS, which is mainly aimed at lightweight small object detection, YOLOv11 is particularly tailored for challenging, high-context tasks. YOLOv11's feature extractor uses C3k2 blocks and a mix of dilated and regular convolutions, which successfully increases the receptive field and allows for more complex motion patterns to be captured. The use of C2PSA self-attention allows for the ability of the model to highlight important posture changes while simultaneously being able to distinguish among co-occurring movements. The neck module further integrates features by combining semantic depth with lower-level features. YOLOv11 also maintains use of a three-scale heads (P3, P4, P5) format, similar to previous versions of YOLO, and stabilizes output through CBS refinement layers. Despite a larger number of parameters (estimated at around 35.8 million), YOLOv11 performed real-time inference with a staggering 8.9 ms per image. Separate ablation studies for its constituent modules were not performed, yet overall findings confirm the benefits of its architecture. Further investigations using controlled ablations (such as disabling C2PSA or replacing C3k2 with standard C3) would help clarify the impact of each module. Nevertheless, performance gains seen in mAP measures, overall generalizability to both dynamic and static scenarios, together with detection correctness, point out the real-world relevance of YOLOv11 for human activity recognition systems.

5 Visualization

All of YOLOv11, YOLOv10, YOLOv9, YOLOv8, ResNet50, and EfficientNet were trained for 50 and 100 epochs. An epoch is a single pass over the entire training set. YOLOv11 continued improving, and the best model was at epoch 100. The same trend appeared for YOLOv10, YOLOv9, and YOLOv8, whereas ResNet50 and EfficientNet were non-converging even for longer training. In terms of training time, YOLOv11 and YOLOv10 took approximately 4 h for 100 epochs, while YOLOv9 and YOLOv8, being lightweight models, took less than 3 h. ResNet50 took approximately 3 h, with EfficientNet, being a lightweight architecture, took less than 2 h for training. Fig. 7 illustrates multiple detected activities in one frame using the YOLOv11-based model.

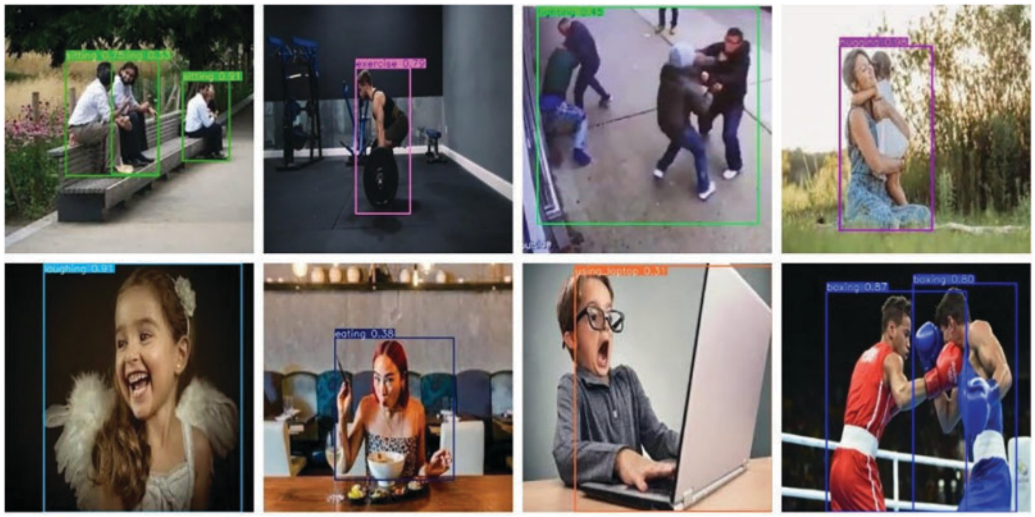


Figure 7: (Continued)

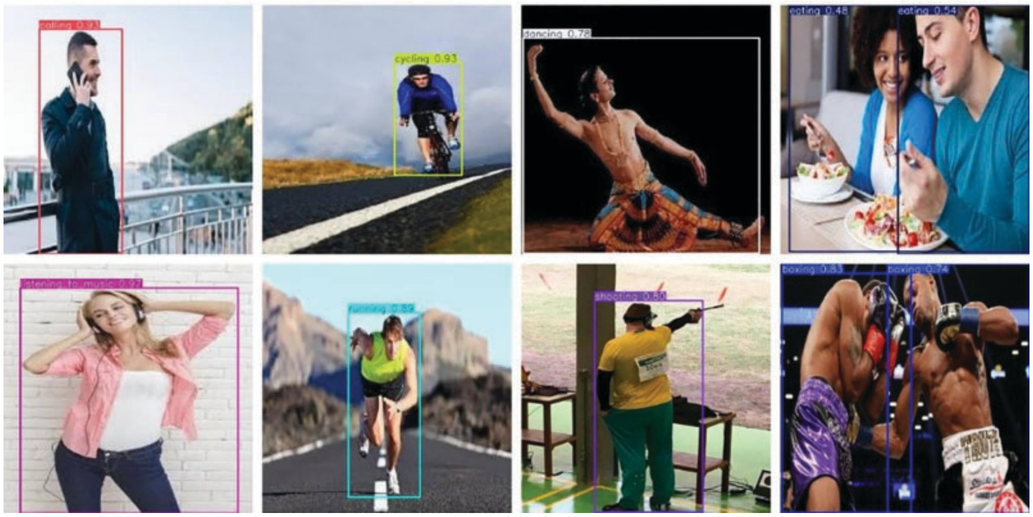


Figure 7: Examples of multiple detected images in one frame using the YOLOv11-based model

Detection outputs were tagged with confidence scores per activity. YOLOv11 performed excellently with high-confidence detections in dynamic and static activities, including running (0.89), dancing (0.78), sitting (0.75, 0.91), sleeping (0.87), eating (0.54, 0.48), boxing (up to 0.87), and hugging (0.98). Improvements were noted across epochs, with YOLOv11 achieving an F1 score of 71%, mAP@0.5 of 74.93%, and mAP@0.5-0.95 of 64.11% at 100 epochs. YOLOv10 performed well with consistent and general performance, particularly for static actions. It resulted in an F1 score of 69%, an mAP@0.5 score of 72.51%, and an mAP@0.5-0.95 score of 61.01% for 100 epochs but had difficulty with discriminations among similar dynamic actions. YOLOv9 performed well for dynamic actions but had difficulty with misclassifications for overlapped static actions. YOLOv8 performed well as a baseline but less accurately with finer discrimination. The ResNet50 and EfficientNet, while performing well on overall tasks, did not generalize for activity detection. They performed with lower F1 values and unstable training patterns and were unreliable for that task. In Fig. 8, Column 1 (YOLOv11): Precise, confident detection of every activity, optimally handling ambiguous cases. Column 2 (YOLOv10): Confident, yet unable to differentiate faint activities. Column 3 (YOLOv9): Average accuracy, misaligned bounding box. Column 4 (YOLOv8): Repeated low confidence for dynamic actions. Column 5 (EfficientNet): Unstable output, low accuracy, and confidence. Column 6 (ResNet50): Weak detection, particularly for dynamic.

Early convergence with stable losses indicated excellent generalization for YOLOv11. YOLOv10 also indicated improvement over epochs, with a trend towards saturation. Overfitting for YOLOv9 using a higher validation loss was indicated. YOLOv8 indicated consistent performance, yet with instability. Overlap with ResNet50 and EfficientNet indicated inefficient learning with unstable training curves. Briefly, YOLOv11 performed consistently better on all measures, followed by YOLOv10 and YOLOv9. YOLOv8 had a decent baseline performance. ResNet50 and EfficientNet, although performing well for general-purpose scenarios, were insufficiently fine-tuned for detailed activity recognition. The experiments on this page validate that architectural improvements, and adequate training epochs are needed for top-end human activity detection.

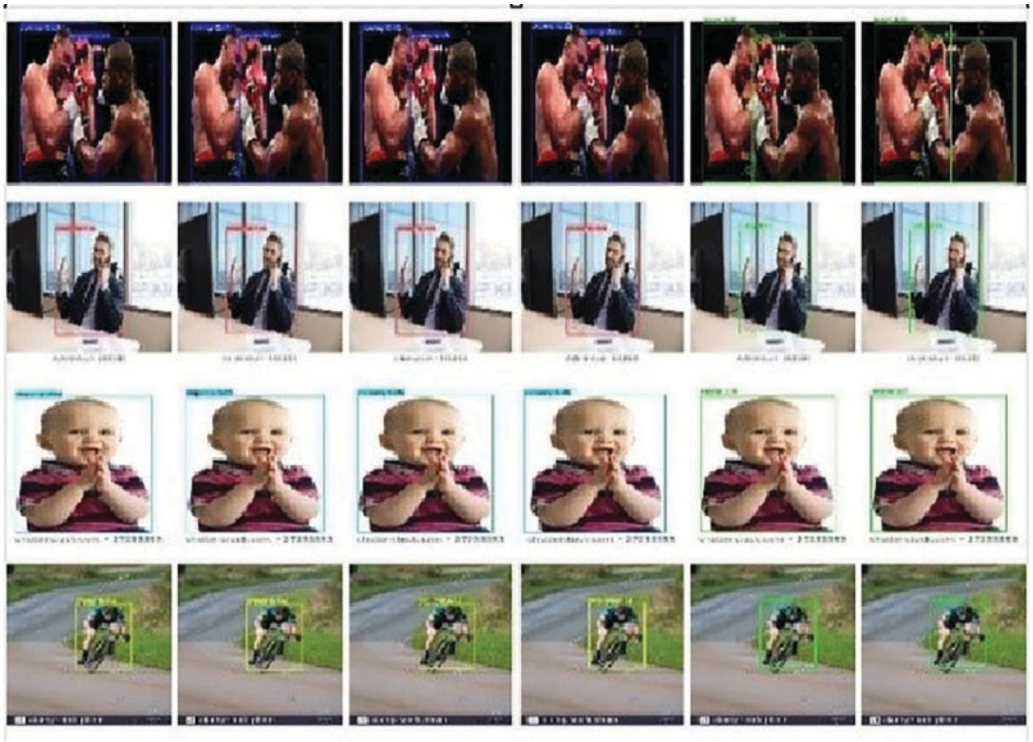


Figure 8: Sample detected images for YOLOv11, YOLOv10, YOLOv9, YOLOv8, EfficientNet, and ResNet50, respectively

6 Conclusion and Future Work

As the demand for human activity understanding intelligent systems grows, this study compared the performance of various models—YOLOv8, YOLOv9, YOLOv10, ResNet50, EfficientNet, and recently introduced YOLOv11—in classifying 19 classes of human activity. The proposed YOLOv11 was best with a high mAP@0.5 of 74. 93% and mAP@0.5:0.95 of 64.11%, consistent accuracy, and recall of both dynamic activities (e.g., running, boxing) and static activities (e.g., sitting, sleeping). It also presented fast inference (8.9 ms/image), which made it appropriate for real-time applications such as surveillance, healthcare, and human-computer interaction. General-purpose models such as ResNet50 and EfficientNet could not cope with the specificity of activity detection, but YOLOv11's improvements, its attention-based, anchor-free design, enabled more accurate and effective performance specific to this field.

Excellent results, despite all, there remain limitations. The dataset, while diverse, was based on publicly released data, and such may not have represented real-world complexity. Specific overlap actions (e.g., sitting and laptop use) were mislabeled, and fine or delicate movements were challenging. To address these, future work can incorporate multimodal data (e.g., using RGB with depth or motion sensors), expand the dataset with more diverse real-world activities, and add complex or culturally specific action classes. Greater optimization of YOLOv11 for mobile and edge deployment will make it more effective in resource-constrained settings. Additionally, integration of self-supervised learning (SSL) methods, such as contrastive learning or image inpainting, can improve generalization and allow the model to perform well under real-world conditions, such as motion blur, occlusions, and varying lighting. These developments will not only make YOLOv11 an extremely high-performing model but also a benchmark for research in the field of robust, flexible, and efficient human action recognition systems for real-world applications.

Acknowledgement: The authors extend their appreciation to the Research Supporting Project, King Saud University, Riyadh, Saudi Arabia, for funding this research.

Funding Statement: This work was supported by King Saud University, Riyadh, Saudi Arabia, under Ongoing Research Funding Program (ORF-2025-951).

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization: Nayeemul Islam Nayeem, Shirin Mahbuba, Sanjida Islam Disha, Md Rifat Hossain Buiyan, Shakila Rahman; Methodology: Nayeemul Islam Nayeem, Shakila Rahman; Software: Nayeemul Islam Nayeem, Md Rifat Hossain Buiyan; Validation: Nayeemul Islam Nayeem, Md Rifat Hossain Buiyan, Shirin Mahbuba, Shakila Rahman; Formal analysis: Nayeemul Islam Nayeem, Md Rifat Hossain Buiyan, Sanjida Islam Disha, Shakila Rahman; Data curation: Nayeemul Islam Nayeem, Md Rifat Hossain Buiyan; Writing: Shirin Mahbuba, Sanjida Islam Disha, Md Rifat Hossain Buiyan; Visualization: Md Rifat Hossain Buiyan, Nayeemul Islam Nayeem, Shakila Rahman; Writing—review and editing: Jia Uddin, M. Abdullah-Al-Wadud; Supervision: Shakila Rahman; Project administration: M. Abdullah-Al-Wadud, Jia Uddin; Funding acquisition: M. Abdullah-Al-Wadud. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data set used for this study can be found at this URL: https://github.com/Krak3n909/YOLOv11_HAR/tree/main (accessed on 24 April 2025).

Ethics Approval: This research did not involve human participants or animals; therefore, ethical approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Anish A, Sharan R, Malini AH, Archana T. Enhancing surveillance systems with YOLO algorithm for real-time object detection and tracking. In: 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS); 2023 Dec 11–13; Pudukkottai, India. p. 1254–7.
2. Waghumbare A, Singh U, Singhal N. DCNN based human activity recognition using micro-doppler signatures. In: 2022 IEEE Bombay Section Signature Conference (IBSSC); 2022 Dec 8–10; Mumbai, India. p. 1–6.
3. Snehitha B, Sreeya RS, Manikandan VM. Human activity detection from still images using deep learning techniques. In: 2021 International Conference on Control, Automation, Power and Signal Processing (CAPS); 2021 Dec 10–12; Jabalpur, India. p. 1–5.
4. Choi B, An W, Kang H. Human action recognition method using YOLO and OpenPose. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC); 2022 Oct 19–21; Jeju Island, Republic of Korea. p. 1786–8.
5. Chen W, Huang H, Peng S, Zhou C, Zhang C. YOLO-face: a real-time face detector. Vis Comput. 2020;37(4):805–13. doi:10.1007/s00371-020-01831-7.
6. Wang C, Yan J. A comprehensive survey of RGB-based and skeleton-based human action recognition. IEEE Access. 2023;11:53880–98. doi:10.1109/access.2023.3282311.
7. Gao D, Wang L. Multi-scale convolution transformer for human activity detection. In: 2022 IEEE 8th International Conference on Computer and Communications (ICCC); 2022 Dec 9–12; Chengdu, China. p. 2171–5.
8. Franco A, Magnani A, Maio D. A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recognit Lett. 2020;131:293–9. doi:10.1016/j.patrec.2020.01.010.
9. Gu F, Chung MH, Chignell M, Valaee S, Zhou B, Liu X. A survey on deep learning for human activity recognition. ACM Comput Surv. 2021;54(8):1–34. doi:10.1145/3472290.
10. Deepthy GS, Peter K, Mathew ET, Bijoy M, Jojo M. Human activity recognition using low cost doppler radar sensor network and CNN. In: 2023 9th International Conference on Smart Computing and Communications (ICSCC); 2023 Aug 17–19; Kochi, Kerala, India. p. 548–53.

11. Ullmann I, Guendel RG, Kruse NC, Fioranelli F, Yarovoy A. Radar-based continuous human activity recognition with multi-label classification. In: 2023 IEEE SENSORS; 2023 Oct 29–Nov 1; Vienna, Austria.
12. Khan IU, Afzal S, Lee JW. Human activity recognition via hybrid deep learning based model. Sensors. 2022;22(1):323. doi:10.3390/s22010323.
13. Liu B, Cai H, Ju Z, Liu H. RGB-D sensing based human action and interaction analysis: a survey. Pattern Recognit. 2019;94:1–12. doi:10.1016/j.patcog.2019.05.020.
14. Li B, Cui W, Wang W, Zhang L, Chen Z, Wu M. Two-stream convolution augmented transformer for human activity recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021 May 19–21; Online. p. 286–93. doi:10.1609/aaai.v35i1.16103.
15. Li Q, Lin W, Li J. Human activity recognition using dynamic representation and matching of skeleton feature sequences from RGB-D images. Signal Process Image Commun. 2018;68:265–72. doi:10.1016/j.image.2018.06.013.
16. Ko MP, Su C, Shie H. Human activity recognition system using angle inclination method and keypoints descriptor network. In: 2024 Conference of Young Researchers in Electrical and Electronic Engineering (ElCon); 2024 Jan 29–31; Saint Petersburg, Russian Federation. p. 235–9.
17. Atikuzzaman M, Rahman TR, Wazed E, Hossain MP, Islam MZ. Human activity recognition system from different poses with CNN. In: 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI); 2020 Dec 19–20; Dhaka, Bangladesh. p. 1–5.
18. Wang M, Zhao Y, Wu Q, Chen G. A YOLO-based method for improper behavior predictions. In: 2023 IEEE International Conference on Contemporary Computing and Communications (InC4); 2023 Apr 21–22; Bangalore, India. p. 1–4.
19. Motwani NP, Soumya S. Human activities detection using deep learning technique-YOLOv8. In: ITM Web of Conferences. Les Ulis, France: EDP Sciences; 2023. Vol. 56. doi:10.1051/itmconf/20235603003.
20. Ha MH. Top-heavy CapsNets based on spatiotemporal non-local for action recognition. J Comput Theor Appl. 2024;2(1):39–50. doi:10.62411/jcta.10551.
21. Choudhury NA, Soni B. Enhanced complex human activity recognition system: a proficient deep learning framework exploiting physiological sensors and feature learning. IEEE Sens Lett. 2023;7(11):1–4. doi:10.1109/lsens.2023.3326126.
22. Rahman S, Rony JH, Uddin J, Samad MA. Real-time obstacle detection with YOLOv8 in a WSN using UAV aerial photography. J Imaging. 2023;9(10):216. doi:10.3390/jimaging9100216.
23. Xu S, Ji Y, Wang G, Jin L, Wang H. GFSPP-YOLO: a light YOLO model based on group fast spatial pyramid pooling. In: 2023 IEEE 11th International Conference on Information, Communication and Networks (ICICN); 2023 Aug 17–20; Xi'an, China. p. 733–8.
24. Shinde S, Kothari A, Gupta V. YOLO based human action recognition and localization. Procedia Comput Sci. 2018;133:831–8. doi:10.1016/j.procs.2018.07.112.
25. Singh T, Vishwakarma DK. A deeply coupled ConvNet for human activity recognition using dynamic and RGB images. Neural Comput Appl. 2020;33(1):469–85. doi:10.1007/s00521-020-05018-y.
26. Sun T, Chen H, Duan X, Lou H, Liu H. Small object detection method based on YOLOv5 improved model. In: 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE); 2022 Sep 23–25; Dalian, China. p. 934–40.
27. Twinkle T, Kaur B, Goel P. Ensembled pretrained convolutional neural network techniques for human activity detection and recognition. In: 2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT); 2024 Aug 29–31; Greater Noida, India. p. 1–6.
28. Mmereki W, Jamisola RS, Mpoeleng D, Petso T. YOLOv3-based human activity recognition as viewed from a moving high-altitude aerial camera. In: 2021 7th International Conference on Automation, Robotics and Applications (ICARA); 2021 Feb 4–6; Prague, Czech Republic. p. 241–6.
29. Agrawal P, Jain G, Shukla S, Gupta S, Kothari D, Jain R, et al. YOLO algorithm implementation for real time object detection and tracking. In: 2022 IEEE Students Conference on Engineering and Systems (SCES); 2022 Jul 1–3; Prayagraj, India. p. 1–6.