Comparative Analysis of YOLOv5/v8/v9 for Object Detection, Tracking, and Human Action Recognition in Combat Sports

Evan Quinn and Niall Corcoran
Technological University of the Shannon: Midlands Midwest, Limerick, Ireland

evan.quinn@tus.ie
niall.corcoran@tus.ie

Abstract: YOLO models are widely used object detectors in computer vision (CV). This study investigates the relative performance of YOLOv5, YOLOv8, and YOLOv9 for object detection, tracking, and human action recognition in combat sports. The models were evaluated using curated datasets encompassing various combat scenarios, athlete movements, and equipment configurations. Pre-processing protocols and augmentation techniques were applied to improve model accuracy and generalizability, including automated orientation correction, image dimension standardisation, contrast enhancement, and methods such as zoom, rotation, shear, and grayscale conversion. The key findings provide insight into the comparative performance of the models across various evaluation metrics, such as precision, recall, and mean average precision. Each model's ability to detect, track, and recognise human actions in dynamic combat sports environments is evaluated. Computational efficiency and real-time performance were assessed as these are important indicators for practical applications in coaching, training, and competitive scoring systems. The findings suggest that YOLOv8 offers the best balance of precision and recall, making it particularly suitable for real-time applications in combat sports analytics. This study contributes to advancing CV technologies in combat sports analytics, with potential implications for improving athletic training methods, facilitating personalised coaching interventions, and enhancing objectivity and consistency in competitive scoring processes in combat sports.

Keywords: YOLOv5, YOLOv8, YOLOv9, Object detection and tracking, Combat sports, Computer vision

1. Introduction

Recent advances in artificial intelligence (AI), computer vision (CV), and decision science are revolutionising many sectors, not least sports, with improvements in several areas including performance analysis, injury prevention, and viewer engagement (Kumar, 2022). The growing field of CV, fuelled by an abundance of visual data, has facilitated athletic performance monitoring and the automation of statistical analysis in sports (Host and Ivašić-Kos, 2022; Jaiswal *et al.*, 2020). For example, Deep Learning (DL) applied to CV has gained interest in sports video understanding, providing richer information to sports consumers and leagues (Van Zandycke *et al.*, 2022). These technologies could be especially effective in combat sports such as boxing, kickboxing, wrestling, and mixed martial arts (MMA), offering precise and unbiased evaluations that address the historical issue of human bias in scoring (e.g., Binter *et al.*, 2024). These disciplines are characterised by one-on-one engagements with specific rulesets and supervised by a referee. Expert judges are heavily relied on to identify athletes objectively, but this process is prone to human error. Furthermore, each combat sport requires a different strategy due to the varying complexities involved in each (Makovsky, 2020). For example, MMA scoring evaluates a wide range of techniques that may be more complex than those found in other combat sports, such as boxing, including strike energy and submission-hold skill (James *et al.*, 2017). New AI-driven technologies offer a fresh lens, promising precision, and unbiased evaluations.

CV has already shown promise in reducing human error in combat sports analysis, such as determining the velocity and impact of punches (Ishac and Eager, 2021) and measuring combat sports athlete performance using wearable sensors (Saponara, 2017). CV has the potential to provide several benefits for combat sports, including scientific training, more effective technical and tactical analysis, and the generation of performance metrics and statistics. Pang *et al.* (2022) highlight the versatility of CV in combat sports, claiming it promotes scientific training, strengthens technical and tactical scrutiny, and generates performance metrics. CV streamlines the analysis, providing trainers and athletes with data-driven insights for improving performance and tactics. Echeverria and Santos (2021) demonstrated this with a study on punch anticipation in karate, offering a comprehensive understanding of martial body movements. The contribution of DL to CV, demonstrated by Van Zandycke *et al.* (2022), further supports this approach. Models built from large datasets, with high-resolution images and detailed annotations, provide comprehensive insights for sports enthusiasts and professionals. According to Pang *et al.* (2022), direct comparison and similarity calculations can be used to evaluate human action attributes.

Using CV techniques in karate has resulted in significant advances in human action recognition (HAR) research for combat sports tracking (Echeverria and Santos, 2021). Convolutional neural networks (CNN) have proven

useful in image recognition and pixel data processing tasks in CV (Patil and Rane, 2021), with potential applications in combat sports. Recent advances in object detection and classification, such as the You Only Look Once (YOLO) architecture, have simplified the process, allowing for real-time performance with high accuracy (Yin *et al.*, 2020). According to Bochkovskiy *et al.* (2020), CNN combined with YOLO has paved the way for object detection and pose estimation in combat sports. HAR is useful for identifying and analysing human movement in videos, but it has been difficult to apply due to the large number of athletes participating in team sports such as football or volleyball (Host and Ivašić-Kos, 2022). Therefore, HAR may be more suitable for combat sports analysis due to the limited number of athletes in a competitive scenario. Accurate HAR is crucial in extracting meaningful information from combat sports videos. Pose detection, a widely used CV task, enables the prediction of human poses, facilitating HAR and action classification (Angelini *et al.*, 2019). Furthermore, the adoption of time-series analysis allows for accurate tracking of identifiers, including pose estimation features, addressing the challenges of long-range context analysis and vanishing gradient problems (Ning *et al.*, 2017). Several studies examine the different YOLO algorithms and compare them for performance and accuracy for real-time single-shot object detection and tracking (e.g., Atik *et al.*, 2022; Jiang *et al.*, 2022; Terven and Cordova-Esparza, 2023).

The primary goal of this study is to investigate the capability of CV for automating scoring in combat sports by comparing automated judgments to human evaluations. Given the intricacies of martial arts movements and the sport-specific rules, crafting a universally embraced scoring algorithm is challenging. It is important to determine the most appropriate architectures for performance and scoring analysis based on CV systems that can categorize athletes based on factors such as attire, stance, and physical characteristics. As part of the main research project, this paper sets out to report on a comparison of three versions of the YOLO model using the same dataset to determine which performs best on the computer hardware. This evaluation will assist with model selection for further research. The focus is to identify the most effective CV architectures for performance and scoring analysis, analyse combat sports footage, and categorise athletes. Additional goals are to determine effective data processing methods, establish best practices for curating labelled combat sports footage, and minimize inaccuracies from low-quality datasets.

1.1 YOLO Models

YOLO is a popular object detection model known for its speed and accuracy (Yin, Li and Fu, 2020). YOLO performs object detection as a regression problem and returns the class probabilities of the detected images. It is a real-time algorithm that detects various objects in images and videos. It was first introduced by Redmon *et al.* (2016) and has had several iterations since then. In this study, YOLOv5, YOLOv8, and YOLOv9 are selected to explore the advancements in the most recent versions and their applicability to combat sports analysis. YOLOv5, implemented in PyTorch and pre-trained on the COCO dataset (Veit *et al.*, 2016), serves as a robust baseline widely adopted for its balance between speed and accuracy Yang *et al.* (2020). It achieves real-time detection at over 140 frames per second (FPS) and has been successfully applied in various industrial applications. YOLOv8 introduces significant enhancements over previous versions, incorporating new elements such as attention modules, self-attention mechanisms, spatial pyramid pooling, and deformable convolutions, which improve detection accuracy and speed while reducing computational demands (Sohan *et al.*, 2024). These advancements make it particularly suitable for combat sports analysis, where precise and efficient detection is crucial. YOLOv9 incorporates new features such as Programmable Gradient Information (PGI) and the Generalized Efficient Layer Aggregation Network (GELAN) to address data loss in deep networks, achieving superior results on the MS COCO dataset (Wang *et al.*, 2024). These innovations can potentially enhance detection accuracy in complex scenarios such as those found in combat sports. YOLOv6 and YOLOv7 are excluded from the analysis for specific reasons. YOLOv6, despite its optimisations for industrial applications, lacks widespread community support and open-source collaboration, making it less suitable for a comparative study aimed at general applicability. YOLOv7, while achieving impressive performance improvements and introducing real-time pose estimation and instance segmentation (Boesch, 2023), has been succeeded by YOLOv8, which builds upon YOLOv7's advancements and offers further enhancements with a more streamlined architecture.

Focusing on YOLOv5, YOLOv8, and YOLOv9 aims to assess the impact of the most recent developments in the YOLO family on object detection, tracking, and HAR in combat sports. This approach allows us to evaluate models representing significant milestones in the evolution of the YOLO architecture and determine their suitability. Some practical applications of YOLO include assembly-line object recognition (Liau and Ryu, 2021), real-time underwater maritime object detection (Yu *et al.*, 2022), facial mask detection (Yang *et al.*, 2020), and diabetic foot ulcer detection (Yap *et al.*, 2021). These examples underscore the effectiveness of YOLO architectures in

different domains, reinforcing the relevance of comparative analysis in combat sports. Table 1 summarises the main features and advancements of YOLO versions v5 to v9.

Table 1: Summary of YOLO Versions v5 to v9

Version	Year	Key Features	Advantages	Limitations
v5	2020	CSPDarknet-backed PyTorch Implementation	High Inference Speed & Easy Deployment	Not Realeased from Original Authors
v6	2022	Efficient Backbone	Optimized for Industrial Applications & High FPS on Specific Hardware	Limited Open Source Adoption
v7	2022	Extended Efficient Layer Aggregation Networks (E-ELAN) Real-time Pose Estimation	Improved Accuracy and Speed State-of-the-Art Performance	Higher Computational Demands Increased Complexity
v8	2023	New Backbone and Neck Architecture Anchor-free Detection Advanced Training Strategies Real-time Pose Estimation	Enhanced Performance Simplified Model Structure Better Generalisation Flexability and Modularity	Still Maturing Community Support with Limited Documentation
v9	2023	GELAN and PGI	Improved Gradient Flow and Superior Benchmark Accuracy	Higher Computational Demands

2. Research Design

The precise identification and analysis of combat athletes were performed using advanced CV techniques, with all models trained on a T4 graphics processing unit (GPU) via Google Colab (NVIDIA, 2018). Image preprocessing techniques such as cropping, flipping, rotation, translation, brightness adjustment, contrast enhancement, colour augmentation, and saturation adjustment were used to improve high-resolution raw images. The study's objectives include tracking multiple elements in combat sports, implementing an algorithm for classifying combat sports actions, and designing an algorithm that uses real-time data for precise scoring and comprehensive performance analysis of athletes. Unlike other methods of HAR, this approach does not use pose estimation. YOLOv5, YOLOv8, and YOLOv9 were selected to evaluate various iterations of the same model, aiming to identify challenges and enhance further research. The accuracy of positive predictions was measured using precision, relevant instances were detected with recall, and average precision across all classes was determined using mean Average Precision (mAP). Real-time performance was measured in frames per second (FPS), with computational efficiency taking precedence in this study due to the inherently efficient nature of YOLO models. Extensive training was conducted with the YOLOv5 model, utilizing an image size of 640, a batch size of 48, and 500 epochs. To accelerate the training process, caching was enabled. The medium model for YOLOv5/v8/v9 is selected by default during the experiments. However, it is worth noting that performance gains can be obtained by selecting the larger or extra-large variants of these models, assuming adequate hardware resources are available.

2.1 Dataset Collection and Pre-Processing

Dataset collection involved curating diverse images from publicly available sources and self-recorded footage to encompass various combat sports scenarios, including boxing, MMA, and kickboxing. A combination of YouTube videos and live recordings was used to ensure variability in athlete movements, equipment configurations, and environmental conditions. The videos featured various athlete actions such as punches, kicks, defensive manoeuvres, and equipment such as gloves and headgear. Frames were extracted from these videos at specific intervals to capture critical moments and a broad spectrum of movements. This approach resulted in a dataset of 5,261 images. To maintain consistency and avoid bias, the dataset was divided into 70% for the training set, 20% for the test set, and 10% for the validation set, as shown in Figure 1. Pre-processing was an essential step in preparing the dataset for training. Automated orientation correction algorithms were implemented to ensure image consistency, rectifying any rotational discrepancies inherent in the raw video frames. All images were resized to a uniform dimension of 640×640 pixels to standardise input across all models, facilitating a fair comparison and efficient batch processing. Contrast enhancement techniques were applied to improve image quality, ensuring that features relevant for detection and recognition were pronounced. Various augmentation methods were used to resolve class imbalances in the dataset. Images were randomly cropped at 0%–15% zoom.

Rotations of -3° to +3°, horizontal shear of ±6°, and vertical shear of ±3° were used to replicate various viewing angles and perspectives. For various illumination, 10% of pictures were grayscaled. Adjustments were performed

to hue (±4°), saturation (±5%), brightness (±3%), and exposure (±9%) to simulate different lighting conditions. Sensor noise and motion blur in video frames were simulated by blurring up to 2 pixels and adding noise up to 0.15%. Three additional outputs per training sample increased dataset diversity and model generalisation. Combat sports data performance increased with pre-trained YOLO model weights and transfer learning from large datasets. Training data augmentation reinforced model overfitting resistance. YOLOv5 and v8 used existing repositories, whereas v9 was adjusted for the framework. Grid search optimised learning rate, batch size, and epochs. Early stopping and learning rate schedulers optimised training time and prevented overfitting, guaranteeing models converged without further computation. This dataset reflects combat sports' dynamic and varied nature by merging YouTube and live frames with considerable pre-processing and augmentation.

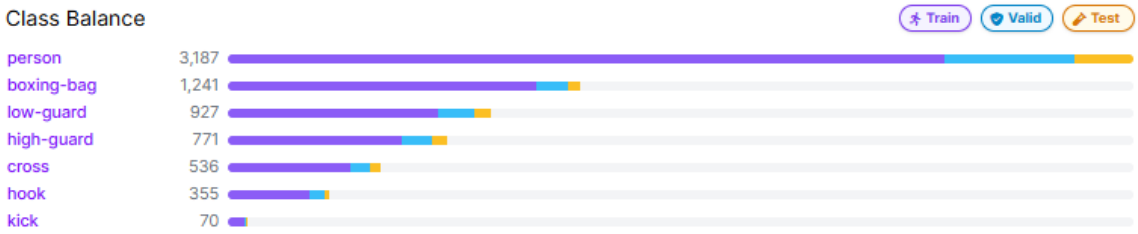

Figure 1: Dataset Classes Balance Breakdown

2.2 Experimentation

The objective is to evaluate the performance of YOLOv5, YOLOv8, and YOLOv9 in detecting objects such as athletes, gloves, and rings by using diverse combat sports datasets that include various scenes and conditions, and then apply standard object detection benchmarks and metrics, including mean Average Precision (mAP), and Precision-Recall metrics. The evaluation compared each model's detection accuracy, ability to handle occlusions, and varying object scales. A multi-object tracking algorithm using detections from each YOLO model was implemented and evaluated. This involved developing a tracking pipeline that integrates these detections within a multi-object tracking framework and evaluating tracking performance using metrics such as accuracy and the number of identity switches. During the evaluation process, particular emphasis was placed on assessing the stability of tracking across both slow-paced and fast-paced sequences, analysing the frequency and causes of identity switches to understand the algorithm's robustness in different conditions. Feature extraction was used to detect objects and extract relevant features for HAR to recognise human actions. The classifier used machine learning (ML) classifiers to analyse the detected features and identify actions. Performance was evaluated using metrics for accuracy, precision, and recall. During the evaluation process, the models were tested on labelled datasets of combat sports actions, analysing performance across different action categories such as punches, kicks, and defensive manoeuvres.

2.3 Computational Efficiency

The computational efficiency of each YOLO model was evaluated by measuring the average inference time per image to determine real-time applicability and memory consumption during model execution. The size of each model was compared to assess storage requirements, and post training tests were performed on various hardware setups, including high-performance GPUs, to understand the impact of the computer hardware on model performance. Computational metrics were recorded and analysed to identify the most efficient model for deployment scenarios. The experimental setup facilitated a comprehensive evaluation of YOLOv5, YOLOv8, and YOLOv9 models across three key areas: object detection, tracking, and HAR. For object detection, the models were assessed on their ability to accurately identify and localise athletes and equipment in different combat sports scenarios using mAP to provide a detailed performance analysis.

2.4 Experimental Setup

All experiments were conducted using Google Colab Pro, which provides access to high-performance GPUs. The hardware and software specifications are shown in Table 2. Additional libraries including NumPy, OpenCV, Matplotlib, and other dependencies as required by each YOLO implementation, were also used.

Table 2: Hardware and Software Specifications

Hardware		Software	Version
GPU	NVIDIA Tesla T4 with 16 GB GDDR6 memory	Ubuntu OS	18.04.6 LTS
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz	Python	3.8.16
RAM	25GB	PyTorch	1.13.1+cu116
		CUDA	11.6
		NVIDIA GPU Drivers	SMI 510.47.03

3. Results

3.1 Model Training Results

Table 3 shows YOLOv5, v8, and v9 training computational performance characteristics. These measurements show model differences in training length, resource use, and final model size. YOLOv8's reasonable model size and low GPU memory suit it for low-GPU situations. YOLOv5, the smallest model, is appropriate for devices with low storage. YOLOv9's memory limits required batch size and epoch modifications, which increased training time and model size.

Table 3: Computational Performance Metrics for YOLO Models

Model	Epochs (Completed)	Training Duration (Hours)	Peak GPU Memory Usage (GB)	Model Size (MB)
YOLOv5	214	1.866	12.7	40.3
YOLOv8	189	2.069	7.28	49.6
YOLOv9	99	4.092	12.9	49

YOLOv5 showed robustness in detecting static objects, while YOLOv8 excelled in scenarios involving rapid movements. YOLOv8's improved stability was particularly evident in sequences with fast-paced actions, improving overall tracking performance. Performance metrics for each model are presented in Tables 4, 5, and 6 respectively, and include accuracy, precision, and recall, providing a comprehensive evaluation of each model's effectiveness in HAR. The computational efficiency of each model was thoroughly tested by measuring inference time, memory usage, and model size. These tests revealed YOLOv5 as the most efficient model for resource-constrained environments, while YOLOv8 balanced performance and computational requirements, making it suitable for real-time applications. The training process included a patience value of 100 for early stopping and halting training early due to no improvements over the last 100 epochs, with the best performance observed at epoch 113. The results summary showed 214 epochs completed over a training duration of 1.866 hours, using 21.5G of memory. Performance metrics for YOLOv5 included an average precision (P) of 0.839 and an average recall (R) of 0.597. Table 4 details the training results for YOLOv5.

Table 4: YOLOv5 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.83	0.894	0.78
cross	500	105	0.82	0.72	0.772	0.481
high-guard	500	192	0.77	0.75	0.76	0.399
hook	500	92	0.65	0.16	0.201	0.117
kick	500	8	1	0	0	0
low-guard	500	213	0.74	0.77	0.725	0.434
person	500	663	0.92	0.96	0.967	0.763

Similar to YOLOv5, early stopping was used for the YOLOv8 training results, saving the best model at epoch 89 after no improvements were seen in the last 100 epochs, using the model yolov8m.pt (medium) with 250 epochs and an image size of 640. The results summary showed 189 epochs completed during a training duration of 2.069 hours, utilising 7.28G of memory. Performance metrics included precision (P) of 0.704, recall (R) of 0.564,

mAP50 of 0.639, and mAP50-95 of 0.451, reflecting its balanced performance across various evaluation criteria. YOLOv8 training results are detailed in Table 5.

Table 5: YOLOv8 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.58	0.888	0.821
cross	500	105	0.78	0.69	0.744	0.489
high-guard	500	192	0.74	0.65	0.71	0.374
hook	500	92	0.29	0.14	0.193	0.0914
kick	500	8	0.53	0.25	0.259	0.146
low-guard	500	213	0.71	0.68	0.709	0.454
person	500	663	0.88	0.96	0.972	0.779

The YOLOv8 model demonstrated strong performance in identifying the 'boxing-bag' and 'person' classes but exhibited lower accuracy in detecting the 'hook' and 'kick' classes, mirroring the results for YOLOv5. Training with yolov8_large or other variants may enhance performance and accuracy, although improving the dataset is important to improving the performance of detections on under-represented classes. YOLOv9 was trained with a batch size of 48, 150 epochs, and an image size of 640, using pre-trained weights gelan-c.pt. The YOLOv9 `gelan-c.pt` model leverages the GELAN to enhance object detection performance, offering efficient and accurate inference capabilities suitable for various computational environments. The training process encountered GPU memory constraints necessitating adjustments in batch size and epochs. The results summary indicated 99 epochs completed over a training duration of 4.092 hours, using 12.9G of memory. Performance metrics included average precision (P) of 0.661, average recall (R) of 0.606, average mAP50 of 0.637, and average mAP50-95 of 0.456. Table 6 details the training results for YOLOv9 and shows strong precision for 'boxing-bag' and 'person' classes but deficiencies in 'kick' and 'hook' detection, suggesting potential gains from dataset augmentation or hyperparameter adjustments.

Table 6: YOLOv9 Training Results

Class	Images	Instances	P	R	mAP50	mAP50-95
boxing-bag	500	309	0.99	0.85	0.915	0.843
cross	500	105	0.78	0.71	0.763	0.516
high-guard	500	192	0.77	0.76	0.784	0.429
hook	500	92	0.39	0.21	0.235	0.116
kick	500	8	0	0	0.0257	0.0103
low-guard	500	213	0.77	0.75	0.769	0.461
person	500	663	0.92	0.96	0.968	0.818

3.2 Model Performance

YOLOv5 and YOLOv8 displayed robust performance with high precision in detecting the 'boxing-bag' and 'person' classes, suggesting that these models are well-suited for tasks involving these objects. However, all models struggled with the 'kick' and 'hook' classes, indicating a need for dataset augmentation or further hyperparameter tuning. YOLOv8 showed an efficient training process with balanced GPU memory utilisation and training time. YOLOv9 encountered memory issues requiring batch size and epochs adjustments, leading to longer training times. Early stopping optimised training durations and prevented overfitting, indicating that better detection of 'hook' and 'kick' requires more representative data. Data augmentation and synthetic data generation could enhance model performance for these underrepresented classes. YOLOv9 maintained high precision but had lower recall for specific classes, indicating areas for improvement. While YOLOv5 excelled in precision and recall for key classes, YOLOv8 offered balanced performance, underscoring the need for targeted enhancements in combat sports detection. Implementing early stopping was important in preventing overfitting and saving training time. Adjusting the patience parameter can further optimise training durations without compromising model performance.

The low performance in detecting certain classes, such as 'hook' and 'kick' across all models, highlights the necessity for targeted improvements, either through enhanced training data or refined model architectures. The dataset has the lowest number of 'kick' and 'hook', which suggests that the dataset could be increased to improve performance. Early stopping significantly optimises training efficiency by mitigating overfitting, emphasising the need for more representative data to improve 'kick' and 'hook' class detection. The results from all three models are displayed in Table 7, where the YOLOv5 model achieves high recall and precision for the terms 'boxing-bag' and 'person,' indicating robust performance across multiple classes. In comparison, YOLOv8 offers a balanced trade-off between precision and recall, with notable improvements in detecting 'cross' and 'high-guard' classes. Meanwhile, YOLOv9 maintains high precision but experiences a decline in recall for specific classes such as 'kick' and 'hook', highlighting the need for dataset augmentation or further tuning to enhance its performance in these areas. Overall, the results indicate that while each model has its strengths, there are specific areas where targeted improvements can significantly enhance their effectiveness for combat sports. Given that the dataset is unbalanced and has potential for improvement, the object detection on the validation batch of the three models performed better than anticipated in Figures 2, 3, and 4. YOLOv9 has superior precision and mAP50 and mAP50-95 for 'person' and 'boxing-bag' classes than v8. High-occlusion sequences test YOLOv9's track consistency. YOLOv9 needs more focused data or architecture improvements due to higher memory utilisation, longer training times, and low 'hook' and 'kick' performance.

Table 7: Comparison of YOLOv5/v8/v9 model's metrics

Model	Images	Instances	P	R	mAP50	mAP50-95
YOLOv5	500	1582	0.84	0.6	0.617	0.425
YOLOv8	500	1582	0.7	0.56	0.639	0.451
YOLOv9	500	1582	0.66	0.61	0.637	0.456

Figure 2: YOLOv5 Prediction Results from Validation Batch

Figure 3: YOLOv8 Prediction Results from Validation Batch

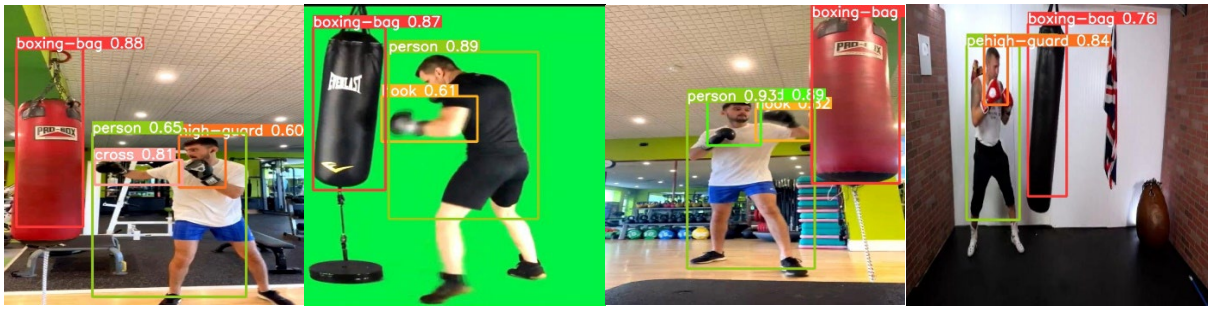

Figure 4: YOLOv9 Prediction Results from Validation Batch

The results provide a detailed breakdown of each model's accuracy in classifying actions, highlighting strengths and areas of misclassification. This pattern of misclassification underscores persistent challenges within dataset and model training processes, particularly for 'hook' and 'kick' actions, necessitating expanded, diverse datasets and enhanced annotation quality for more precise model learning. Addressing these issues is essential for improving model performance, reducing confusion, and enhancing accuracy in recognising complex actions in real-world scenarios. Figure 5 displays scenes from a sample video that contained some frames used to train the model.

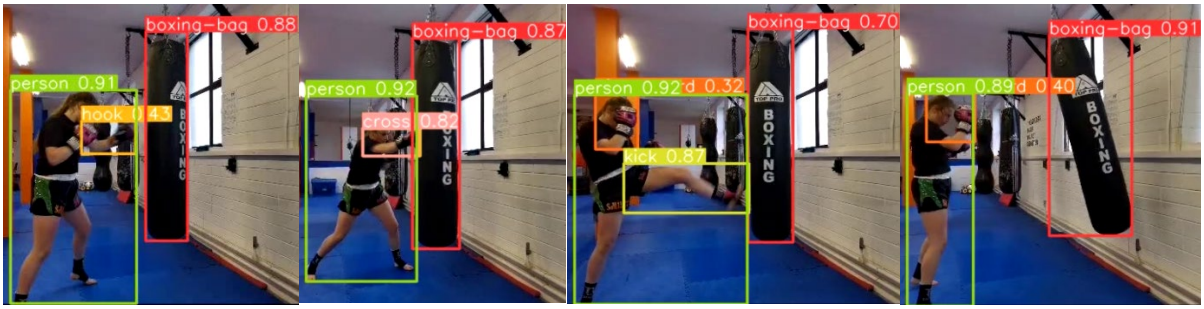

Figure 5: YOLOv8 Results from Video Sample - Hook, Cross, Kick, High Guard, Person, and Boxing-Bag Classes

4. Discussion

Comparing YOLOv5, v8, and v9 in combat sports shows their strengths and weaknesses. Precision and recall are YOLOv5's strengths in 'person' and 'boxing-bag' detection and tracking. Memory utilization and training duration efficiency make it suitable for resource-constrained applications. It misclassifies 'hook' and 'kick' actions, requiring better training data or model optimisation. The balanced YOLOv8 method provides moderate class precision and recall. It identifies 'cross' and 'high-guard,' enhancing tracking stability during rapid movements. Despite these benefits, issues with the 'hook' and 'kick' classes, similar to YOLOv5, suggest dataset limits rather than model architecture problems. YOLOv9 stresses precision but has unpredictable recall, especially for complex categories such as 'hook' and 'kick.' This sophisticated architecture uses GELAN and PGI, resulting in high memory usage and long training times. Improvements in detection accuracy may need additional processing resources, making YOLOv9 more suitable for hardware-capable applications. Dataset class imbalance is seen in all models' inability to detect 'hook' and 'kick' actions. Models cannot learn and generalise these categories without these activities in the training data. This suggests that dataset quality and balance in DL model training is necessary for specific applications. Temporal analysis methods for object detection and tracking may improve combat sports HAR. Transformers and long short-term memory (LSTM) networks improve dynamic action understanding by efficiently managing sequential data and temporal dependencies. These advanced algorithms and YOLO's real-time object identification can improve HAR systems' accuracy and reliability. Coaching and combat sports scoring require quick decisions and feedback, which matches their fast and precise video frame processing.

4.1 Limitations and Future Work

The dataset's class imbalance, especially the under-representation of 'hook' and 'kick' activities, limits this study. The models' poor class detection may have been due to this imbalance. Data augmentation or video of under-represented actions should balance the dataset. Because of its processing needs, YOLOv9 may not be suited for

all applications. This can be solved via model optimisation or hardware acceleration. Future studies could combine temporal analysis with object detection algorithms to improve HAR, possibly using time series analysis.

5. Conclusions

This study examines combat sports object detection, tracking, and HAR using the YOLOv5, YOLOv8, and YOLOv9 object detection systems. A comparative analysis of the three models selected YOLOv8 due to its balanced precision and recall across classes, consistent object identification, and robust performance measures. Its capacity to improve tracking stability during quick movements makes it ideal for coaching, training, and competitive scoring. YOLOv5 accurately detects 'person' and 'boxing-bag' classes but not 'hook' or 'kick' actions. Its training duration and memory economy make it suitable for resource-constrained applications. YOLOv9's precision and superior architecture may increase detection rates. However, it demands more processing resources and sometimes fails to keep consistent tracks in high-occlusion or overlapping action sequences, limiting its applicability. The results show that YOLO models effectively detect and track combat sports object identification, emphasising the need for model selection based on application needs. The difficulty of models detecting specific action classes suggests dataset restrictions must be addressed.

Additional data must be gathered and enhanced to represent under-represented categories, such as ‘hook’ and ‘kick' to improve detection accuracy and efficiency. Enhancing hyperparameters and model architectures should elevate performance in difficult classes without augmenting computational demands. Integrating temporal analysis for HAR is enhanced when YOLO models are combined with temporal networks to capture the temporal dependencies of action sequences. These advancements can enhance HAR systems and the identification of objects in combat sports, thereby improving sports analytics and technology-driven coaching. This study aims to enhance the understanding of CV and AI technologies for sports analytics by examining cutting-edge object detection models within the specific domain of combat sports. The findings can be used to improve training methodologies, monitor athletes' movements and actions, tailor coaching interventions, and increase the objectivity and consistency of competitive scoring.

References

Angelini, F., Fu, Z., Long, Y., Shao, L. and Naqvi, S. M. (2019) '2D Pose-Based Real-Time Human Action Recognition with Occlusion-Handling', *IEEE Transactions on Multimedia,* 22(6), pp. 1433-1446.

Atik, M. E., Duran, Z. and Özgünlük, R. (2022) 'Comparison of YOLO versions for Object Detection from Aerial Images', *International Journal of Environment and Geoinformatics,* 9(2), pp. 87-93.

Binter, J., Říha, D., Boschetti, S., Prossinger-Beck, V., Stachoň, M. and Prossinger, H. (2024) 'Navigating Between Human and Machine-Based Evaluation: Judgment and Objectivity in Economic Games Exemplified in the Analysis of MMA Fights': Springer Nature Switzerland, pp. 235-246.

Bochkovskiy, A., Wang, C. Y. and Liao, H. Y. M. (2020) 'Yolov4: Optimal Speed and Accuracy of Object Detection', *arXiv preprint,* 1(arXiv:2004.10934), pp. 1-17.

Boesch, G. (2023) *YOLOv7: A Powerful Object Detection Algorithm (2024 Guide)*: viso.ai. Available at: https://viso.ai/deep-learning/yolov7-guide/ 2024).

Echeverria, J. and Santos, O. C. 'Punch Anticipation in a Karate Combat with Computer Vision', *UMAP '21: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, Utrecht, Netherlands, 21-25 June: ACM, pp. 61–67.

Host, K. and Ivašić-Kos, M. (2022) 'An Overview of Human Action Recognition in Sports Based on Computer Vision', *Heliyon,* 8(6), pp. 1-25.

Ishac, K. and Eager, D. (2021) 'Evaluating Martial Arts Punching Kinematics Using a Vision and Inertial Sensing System', *Sensors,* 21(6), pp. 1-25.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D. and Makedon, F. (2020) 'A Survey on Contrastive Self-Supervised Learning', *Technologies,* 9(1), pp. 1-22.

James, L. P., Robertson, S., Haff, G. G., Beckman, E. M. and Kelly, V. G. (2017) 'Identifying the Performance Characteristics of a Winning Outcome in Elite Mixed Martial Arts Competition', *Journal of Science and Medicine in Sport,* 20(3), pp. 296-301.

Jiang, P., Ergu, D., Liu, F., Cai, Y. and Ma, B. (2022) 'A Review of Yolo Algorithm Developments', *Procedia computer science,* 199(1), pp. 1066-1073.

Kumar, A. (2022) 'Gamification in Training with Next Generation AI-Virtual Reality, Animation Design and Immersive Technology', *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1-14.

Liau, Y. Y. and Ryu, K. (2021) 'Status Recognition Using Pre-Trained YOLOv5 for Sustainable Human-Robot Collaboration (HRC) System in Mold Assembly', *Sustainability,* 13(21), pp. 1-15.

Makovsky, Z. (2020) 'The 10-point system - a must? A critique of MMA’s scoring system', *The Fight Site*. Available at: https://www.thefight-site.com/home/10-point-system-critique-scoring-mma 2024].

Ning, G., Zhang, Z., Huang, C., Ren, X., Wang, H., Cai, C. and He, Z. 'Spatially supervised recurrent convolutional neural networks for visual object tracking'. *2017 IEEE international symposium on circuits and systems (ISCAS)*, Baltimore, MD, USA, 28-31 May: IEEE, pp. 1-4.

NVIDIA (2018) *NVIDIA Tesla T4 GPU Architecture*: NVIDIA. Available at: https://www.nvidia.com/en-us/data-center/tesla-t4/.

Pang, Y., Wang, Q., Zhang, C., Wang, M. and Wang, Y. 'Analysis of Computer Vision Applied in Martial Arts', *2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Guangzhou, China, 14-16 January. https://ieeexplore.ieee.org/xpl/conhome/9712655/proceeding: IEEE, pp. 1-6.

Patil, A. and Rane, M. 'Convolutional Neural Networks: An Overview and its Applications in Pattern Recognition', *Information and Communication Technology for Intelligent Systems*, Online, May 15-16. Singapore: Springer Nature, pp. 21-30.

Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. 'You Only Look Once: Unified, Real-Time Object Detection', *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 1 January: CVPR, pp. 779-788.

Saponara, S. (2017) 'Wearable Biometric Performance Measurement System for Combat Sports', *IEEE Transactions on Instrumentation and Measurement,* 66(10), pp. 1-11.

Sohan, M., Sai Ram, T., Reddy, R. and Venkata, C. 'A review on yolov8 and its advancements'. *International Conference on Data Intelligence and Cognitive Informatics*: Springer, pp. 529-545.

Terven, J. and Cordova-Esparza, D. (2023) 'A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond', *arXiv preprint arXiv:2304.00501,* 1(1), pp. 1-34.

Van Zandycke, G., Somers, V., Istasse, M., Del Don, C. and Zambrano, D. (2022) 'DeepSportradar-v1: Computer Vision Dataset for Sports Understanding with High Quality Annotations', *MMSports '22: 5th International ACM Workshop on Multimedia Content Analysis in Sports,* 5(5), pp. 1-8.

Veit, A., Matera, T., Neumann, L., Matas, J. and Belongie, S. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140.* IEEE Xplore.

Wang, C.-Y., Yeh, I.-H. and Liao, H.-Y. M. (2024) 'Yolov9: Learning What you Want to Learn Using Programmable Gradient Information', *arXiv preprint arXiv:2402.13616*.

Yang, G., Feng, W., Jin, J., Lei, Q., Li, X., Gui, G. and Wang, W. 'Face mask recognition system with YOLOV5 based on image recognition'. *IEEE 6th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 11-14 December: IEEE, pp. 1398-1404.

Yap, M. H., Hachiuma, R., Alavi, A., Brüngel, R., Cassidy, B., Goyal, M., Zhu, H., Rückert, J., Olshansky, M. and Huang, X. (2021) 'Deep learning in diabetic foot ulcers detection: a comprehensive evaluation', *Computers in Biology and Medicine,* 135(104596), pp. 1-17.

Yin, Y., Li, H. and Fu, W. (2020) 'Faster-YOLO: An accurate and faster object detection method', *Digital Signal Processing,* 102(1), pp. 1-11.

Yu, S., Wu, G., Gu, C. and Fathy, M. E. 'TDT: Teaching Detectors to Track without Fully Annotated Videos', *CVPR 2022 Workshop: Workshop on Learning with Limited Labelled Data for Image and Video Understanding (L3D-IVU)*, New Orleans, Louisiana, USA, 21 - 24 June: IEEE, pp. 3940-3950.