Advancing functional reach assessments: a comparison of YOLOv8 human pose estimation and 3D motion capture in a young healthy cohort

AULTON, Cavan, CHIU, Chuang-Yuan and CHIOU, Shin-Yi

Journal Pre-proof

Advancing functional reach assessments: a comparison of YOLOv8 human pose estimation and 3D motion capture in a young healthy cohort

Cavan Aulton, Chuang-Yuan Chiu, Shin-Yi Chiou

PII: S3051-1313(26)00006-3

DOI: https://doi.org/10.1016/j.jbmo.2026.100007

Reference: JBMO 100007

To appear in:

Accepted date: 28 May 2026

Please cite this article as: C. Aulton, C.-Y. Chiu and S.-Y. Chiou, Advancing functional reach assessments: a comparison of YOLOv8 human pose estimation and 3D motion capture in a young healthy cohort, (2024), https://doi.org/10.1016/j.jbmo.2026.100007

Advancing Functional Reach Assessments: A Comparison of YOLOv8 Human Pose Estimation and 3D Motion Capture in a young healthy cohort.

Cavan Aulton[1*], Chuang-Yuan Chiu[1], Shin-Yi Chiou[2]

[1]Sports Engineering Research Group (SERG), College of Health Wellbeing and Life Sciences, Sheffield Hallam University, Sheffield, UK

[2]School of Sport, Exercise and Rehabilitation Sciences, College of Life and Environmental Sciences, University of Birmingham, Birmingham, UK

Running title: Comparing YOLOv8 and 3D motion capture during functional reach testing.

***co-corresponding authors:**

Mr. Cavan Aulton https://orcid.org/0000-0002-5766-4997

Email: c.aulton@shu.ac.uk

Dr Chuang-Yuan Chiu https://orcid.org/0000-0002-6512-0084

Email: c.chiu@shu.ac.uk

Dr Shin-Yi Chiou https://orcid.org/0000-0002-4200-5243

Email: s.chiou@bham.ac.uk

Abstract

Accurate assessment of dynamic sitting balance is essential in rehabilitation. Traditional 3D motion capture systems provide precise measurements but are costly and impractical for daily clinical use. Human Pose Estimation (HPE) offers a low-cost and less time-consuming alternative to clinical assessments for Spinal Cord Injuries compared to traditional measurement techniques. Sixteen healthy adults performed forward and lateral Modified Functional Reach Tests across two sessions. Movements were recorded using an 8-camera 3D motion capture system and two 2D cameras. YOLOv8n (lightweight version) and YOLOv8x (heavy version) HPE models were applied to estimate reach distances. Reliability was assessed using intraclass correlation coefficients (ICC), standard error of measurement (SEM), and coefficient of variation (COV). Validity was evaluated against 3D motion capture using repeated measures correlation and Bland–Altman analysis. YOLOv8n showed excellent reliability and strong concurrent validity for forward reaching (ICC ≥ 0.98; rm-R^2 = 0.98), while YOLOv8x performed better in the lateral reaching task (ICC ≤ 0.35; rm-R^2 = 0.73). Findings suggest both models underestimated lateral reach distances compared to 3D motion capture but a practical and reliable solution for forward reach assessment and shows potential for clinical application in resource-limited settings. However, reduced reliability and spatial agreement in lateral reaching highlights the need for further refinement and validation in clinical populations. This study provides a foundation for integrating deep learning-based pose estimation into rehabilitation practice, enabling accessible, markerless motion analysis for dynamic balance evaluation.

Key Words: Spinal Cord Injury, Human Pose Estimation, Rehabilitation Technology, Biomechanics, Deep Learning.

1 Introduction

The Modified Functional Reach Test reliably assesses dynamic sitting balance in individuals unable to stand, including those with stroke, Parkinson's Disease, and spinal cord injury (SCI) (Schenkman et al., 1997; Lynch et al., 1998; Katz-Leurer et al., 2009). Balance is crucial for daily function because greater reach distance is associated with faster Timed Up and Go completion (Marchesi et al., 2021) and superior postural control during upper-limb movement (Chiou and Strutton, 2020). Clinically, the test relies on a wall-mounted yardstick and manual assessment which is practical but highly operator-dependent method susceptible to parallax errors (Lee et al., 2022). This manual assessment approach fails to generate the objective, digitised records required to precisely monitor longitudinal changes in patient performance.

Markerless motion capture has emerged as a versatile, cost-effective alternative to traditional marker-based systems, enabling objective kinematic assessment of the upper limbs in both clinical and real-world environments (Francia et al., 2026; Hansen et al., 2024). While these systems demonstrate high reliability in tracking temporal movement patterns and elbow or wrist joint angles, research indicates they may still face challenges such as systematic underestimation of maximum shoulder flexion and sensitivity to joint occlusions (Francia et al., 2026). By replacing the yardstick with a standard camera, clinicians can eliminate subjective errors, reduce assessment time, and instantly digitise data. Digitising and automating this fundamental clinical metric using a two-Dimensional (2D) markerless approach offers significant translational impact. This provides a highly scalable solution for

telerehabilitation and remote monitoring, environments where traditional 3D systems or multi-camera arrays are impossible to deploy.

The gold standard approach to digitally measure reaching distance is using a three-dimensional (3D) optical motion capture system which provides highly reliable and accurate measurements (Scataglini et al., 2024). However, these systems can be prohibitively expensive (ranging from £10,000 to £100,000), requires dedicated laboratory space, and is time-consuming in data acquisition, making it unsuitable for routine clinical environments or telerehabilitation (Scataglini et al., 2024). Moreover, 3D motion capture requires expertise in morphological landmarks to accurately place markers on patients, which may cause errors in measurement outcomes. 2D video-based motion capture systems have emerged as a low-cost, scalable alternative for movement analysis in sports and rehabilitation (Garrido-Castro et al., 2012; Colyer et al., 2018; Reinking et al., 2018). However, its validity compared to 3D systems varies across joints and planes of movement. Studies have reported lower validity in the frontal and transverse planes and in distal joints, such as the ankle, while validity was generally higher in the more proximal joints like the hip and knee (Schurr et al., 2017; Reinking et al., 2018; Leporace et al., 2023).

Human Pose Estimation (HPE) is a method of estimating the position of different body parts from visual information sources, using computer vision and deep learning algorithms to identify the spatial location of joints and bones from 2D images or videos without the need for markers (Boualia and Essoukri Ben Amara 2019; Zhang et al., 2020; Zheng et al., 2022). Recent advancements in this markerless technology have been increasingly applied to upper-limb kinematics to bypass the limitations of traditional systems, such as extensive preparation, marker adhesives, skin motion artifacts, and tight clothing requirements. HPE can be

performed retrospectively from video data and only requires a single camera, making it a more accessible, cost-effective, and efficient alternative to traditional motion analysis methods. These advantages highlight the potential of HPE to be a feasible method for clinical use in quantifying performance of the functional reach test. Recent literature has validated advanced HPE systems against gold-standard marker-based systems for upper extremity clinical assessments, such as the Box and Blocks Test. These studies demonstrate that markerless tracking can achieve good to excellent between-session agreement and clinically acceptable joint angle deviations of less than 6°. However, while highly accurate, these advanced markerless solutions still require complex, multi-camera arrays (e.g., 8-camera setups) and specialised software to generate 3D pose estimations (Hansen et al., 2024). Despite these advancements in multi-camera upper-limb kinematics, the clinical impact of these tools remains limited by their high cost and impracticality for daily use. Also, to our knowledge, no studies have evaluated the reliability or concurrent validity of 2D versus 3D systems specifically for wrist joint displacement. Therefore, the deployment of a single camera 2D HPE is important to quantify absolute spatial displacement (i.e., reach distance) but remains underexplored.

The You Only Look Once (YOLO) model is an open-source computer vision library which can complete different tasks from images and videos, including human detection and HPE (Terven et al., 2023). YOLOv8 models can identify joint locations of humans from single images or videos without manual annotation, using GPU acceleration and CPU multi-threading to achieve good results in real-time using consumer-grade hardware, making it suitable for health setting applications. Prior research has documented the applications of YOLOv8 in the clinical context by analysing fundus images to improve early ophthalmic disease classification (Khalaf and Abdulateef 2024) and analysing MRI images to identify signs of abnormalities

related to brain tumours, improving early detection of tumours (Hashemi et al., 2024). Moreover, Alruwaili et al., (2024) demonstrated how the YOLOv8 models can improve the detection of individuals with disabilities (i.e., detection of individuals in wheelchairs or walking frames), allowing services and support to be tailored to individual needs. Finally, regarding accuracy, Dong et al. (2024) demonstrated that a modified version of YOLOv8 achieved an average precision (AP) of 73.7% and 74.6% on the Microsoft Common Objects in Context (COCO) Keypoint and CrowPose datasets, respectively (Lin et al 2014). These datasets used in previous studies applying the YOLOv8 model are independent and compatible with other models that share the same keypoint output format as YOLOv8, such as Mask R-CNN, HRNet, OpenPose, HigherHRNet, and OpenPifPaf, all of which have been benchmarked on the COCO dataset (Cheng et al., 2020; Lin et al., 2014). Among these, AlphaPose is architecturally closest to YOLOv8 because it leverages a YOLO-based person detector before running a dedicated single-person pose estimator, mirroring YOLOv8's top-down pipeline (Fang et al., 2023). These previous studies demonstrate how YOLOv8 could be leveraged to accurately improve crucial aspects of the care sector, enhancing patient outcomes for a variety of medical needs.

YOLOv8 has several iterations of its model, the primary difference between YOLOv8n (nano) and YOLOv8x (extra-large) lies in their model size, complexity, and intended use cases. YOLOv8n is the smallest and most lightweight variant and is designed for applications requiring high-speed inference on devices with limited computational resources (i.e., mobile phones). YOLOv8n has fewer parameters and lower computational cost, making it ideal for real-time applications with lower accuracy requirements. On the other hand, YOLOv8x is the largest and most robust variant, boasting significantly more parameters and layers, leading to higher accuracy and performance in object detection tasks. But this increase in model

complexity makes YOLOv8x computationally heavier, requiring more powerful hardware such as GPUs for efficient operation.

Existing studies using HPE focus primarily on counting repetitions, timing-based measures (such as sit-to-stand tasks), or joint angle calculations (Wallmann., 2023; Rana., 2024), but using HPE specifically to quantify reach distance remains unexplored in the current research landscape. By validating reach distance estimation using HPE against gold-standard 3D motion capture, the findings may offer a foundation for using simple, camera-based methods in clinical or remote rehabilitation settings where traditional systems may not be feasible. Testing on healthy individuals provides a controlled baseline to evaluate the accuracy and reliability of pose estimation models before extending their application to clinical populations. Therefore, the aim of this study was to evaluate the reliability and concurrent validity of human pose estimation for forward and lateral reaching using the YOLOv8n and YOLOv8x models in healthy adults. We hypothesised that the YOLOv8-based human pose estimation models would demonstrate high reliability (intra- and inter-session) and good concurrent validity when compared to gold-standard 3D marker-based motion capture for measuring forward and lateral reach distances during the Modified Functional Reach Test in healthy adults.

2 Methods

2.1 Participants

The study was approved by University of Birmingham Research Ethics Committee (ERN_20-1453) and performed in accordance with the Declaration of Helsinki. Recruitment was carried out by students at the University of Birmingham and all participants provided written informed consent. Sixteen healthy young adults (age: 21 ± 4 years, height: 172.58 ± 8.33cm, arm length: 55.75 ± 4.03cm, weight: 72.55 ± 18.43kg; 7 males and 9 females) took part in the study. Participants were excluded if they had a history of musculoskeletal abnormalities in the upper extremity, back muscles, or axial skeleton (e.g., scoliosis or low back pain). We selected this sample size as it is like the participant group sizes used within this area of research (Field-Fote and Ray, 2010; Reinking et al., 2018).

2.2 Experimental Procedures

All participants attended the laboratory on two occasions (between-session interval: 7.5 ± 1.2 days) at approximately the same time of day. The participants' height and mass were measured prior to performing forward and lateral reaching in each session. Tasks were performed in a pseudorandom order with all participants completing the same tasks in different orders. Participants were seated upright on a chair (seat height: 48cm), with the trunk unsupported, hips and knees positioned at 90 degrees flexion, feet positioned flat on the floor and arms by the sides (the start position). For consistency, participants were asked to flex the right arm to 90 degrees and reach forward or to the right as far as possible five times (similar procedure can be found in Baltaci et al., 2003). A total of 16 participants each completed five trials of both forward and lateral reaching tasks across two sessions, yielding

an expected 320 trials (160 forward-reaching and 160 lateral-reaching trials). Two forward-reaching sessions from one participant were excluded due to a data collection error, and one lateral-reaching trial from a different participant was removed due to a technical issue with the video recording.

2.3 Reach distance from motion capture

Movements were recorded using an 8-camera 3D optical motion capture system (Smart DX 6000, BTS Bioengineering Corp, Quincy, MA, USA) operating at 250 Hz. Reflective markers were placed bilaterally on the ulnar styloid process to measure reaching distance (Figure 1) (Field-Fote and Ray, 2010), each participant completed two data collection sessions on different days resulting in 32 forward reaching and lateral reaching trials. During forward reaching trials the x axis was used to measure reach distance and the Z axis was used for lateral reaching trial as the motion capture axes are fixed to the room, but YOLOv8 models are camera orientated so the X axis was used for both forward and lateral reach trials.

****Figure 1 Here****

Figure 1 Close up schematic of the marker placement on the ulnar styloid.

2.4 Reach distance from human pose estimation

YOLOV8n and YOLOV8x were applied to identify the joint locations in the video so no training processes are needed for this study. YOLOV8 models are pre-trained on data that do

not contain retroreflective markers, the model likely identified joint locations rather than reflective markers in this study. Nevertheless, the influence of the reflective markers on the YOLOv8 models remains to be tested. Comparing models like YOLOv8n and YOLOv8x is essential because it helps developers choose the most suitable model based on specific application requirements, balancing factors such as accuracy, speed, and computational efficiency to optimise performance in real-world scenarios. This study compared the YOLOv8x and YOLOv8n models because they represent the two extremes of the YOLOv8 family.

Two cameras were used to capture 2D videos of the movement and sampled at a frequency of 25 Hz. The cameras were placed on a tripod at a constant height of 0.90 m from the floor and a constant distance of 2 m from the participant. One camera was placed in the sagittal plane for capturing forward reaching movement, and the other camera was placed in the frontal plane to capture lateral reaching movement; the position of the cameras remained the same throughout the session. The YOLOv8 model implemented by Ultralytics (Ultralytics 8.3.29) in Python (Python 3.11.6) was applied for HPE to identify the joint centres automatically. After automatic HPE, the first frame of each video was processed manually to identify the target person in the video. A person selector plot was created to ensure the model correctly inferred the participant in cases where multiple individuals were present in the video data and the researcher had to press on the image of the person in frame YOLOv8 should run on. It must be stated however the model itself automatically detects all the people in frames; the researcher simply had to select which person they wanted the model to run on because YOLOv8 can run on multiple people in frames at once. The labelled marker data was automatically exported to a .csv file to allow further analysis. Then a Savitsky–Golay filter was applied to the raw pose estimation data effectively smoothing the signal while preserving critical features, such as peaks and valleys, which is essential for accurate biomechanical

analysis (Crenna et al., 2021). Specifically, this filtering was implemented using Python's SciPy library, with a window length of 21 data points and a polynomial order of 3 (Crenna et al., 2021). These parameters were chosen to balance effective noise reduction and signal fidelity. We consider the motion capture data as the ground truth, as it relates to tracking a fixed point and is considered the gold standard of motion capture (Jakob et al. 2021). Unlike the ground truth data, HPE data needs to be scaled to the real world. Therefore, a force plate (400 x 600mm, BTS P6000, BTS Bioengineering Corp, Quincy, MA, USA) visible in the videos was used as a calibration measure to adjust the HPE data to real-world scale. Inaccurate 2D calibration can cause pose estimation errors by distorting the transformation from 2D image points, which can be mitigated by using accurate and robust calibration patterns such as parallel perspective to the movement (Henrichsen, 2000).

During data collection participants were instructed to raise their arm, pause, and then initiate the reach. This was to aid in the detection of a start and end point of the movement and accurately quantify reach distance. To address concerns regarding operator dependency, we developed and validated a velocity-based automated onset detection algorithm applying kinematic thresholds consistent with published approaches (Marchesi et al., 2021). Automated onset detection used a velocity-based kinematic threshold approach. Wrist position was differentiated and smoothed (Savitzky-Golay filter) to obtain velocity profiles. Trial boundaries were defined by global peak detection with valley-based segmentation. Within each trial, movement onset was identified as the first frame exceeding a 50 mm/s velocity threshold following the characteristic low velocity pause at approximately 90° shoulder flexion. This was a process of trial and error that tried numerous thresholds to find the best result. Ultimately, comparison against manually identified onset frames across all trials revealed a mean absolute difference of 9.8 frames (SD=12.7) for forward reaching and

15.6 frames (SD=23.1) for lateral reaching. While agreement was acceptable for forward reaching (median=8 frames, 320ms), lateral reaching showed substantially higher variability (median=10 frames, max=141 frames), with large errors in participants exhibiting atypical approach kinematics. Future research should therefore develop and train a deep learning model to detect key time points in this functional sit and reach test to remove any further manual operator involvement.

Given that starting frame error propagates directly into reach distance calculations, and that automated detection proved insufficiently reliable for lateral reaching, manual identification was retained. The validated automated algorithm, full trial-level comparison data, and the investigator's frame selection reliability results are provided in Appendix 2. Manual identification was utilised to ensure more representative initiation of the forward or lateral reach was captured, starting frames were visually assessed and selected within the centre of the plateau before each peak (see vertical lines on Figure 2). To minimise inter-rater variability, the same operator conducted all frame selections and the reliability results of the operator can be found in Appendix 3. Due to an unforeseen hardware trigger malfunction during data collection, the 2D video cameras and 3D motion capture system lacked a reliable hardware sync pulse and were not temporally synchronised. Therefore, given the additional challenges with automated onset detection, starting frames had to be identified manually and independently for both the ground truth and HPE data. A bespoke Python script was then created to find the peak reach distance in each trial automatically. A dynamic threshold (i.e., 0.95 * maximum) was applied in the script to avoid identifying multiple peaks within the same test (Figure 2), this was done to find the maximum value of each trial rather than the code detecting multiple highest values in the same trial which was common if this threshold wasn't implemented. To calculate reach distance, the X value of the right wrist at the start frame was

subtracted from the X value of the right wrist at the peak frame which is automatically done within a custom python script. The full data collection pipeline is in Figure 3.

****Figure 2 Here****

Figure 2 shows the trace of the right wrist in the x axis as collected by the YOLOv8x model during a forward reaching trial with the dynamic threshold, starting frames and peaks annotated.

****Figure 3 Here****

Figure 3 Flowchart of the YOLOv8-based pose estimation and analysis pipeline, from video capture to reach distance calculation

2.5 Statistical Analysis

Statistical analysis was performed in R Studio (R Studio version 4.3.3), the purpose of this analysis is to evaluate the accuracy and reliability of two YOLOv8 models (V8n and V8x) compared to ground truth motion capture data in the context of forward and lateral reaching sit-and-reach trials. Prior to conducting parametric tests, the assumption of normality was assessed and verified using the Shapiro-Wilk test. A repeated measures correlation analysis was conducted using both all five peak reach distances and the 3 largest peaks per participant (Bakdash & Marusich, 2017). Analysing all five peaks offers a comprehensive view of consistency and variability, while focusing on the top 3 targets' maximum performance, reducing the influence of lower or inconsistent values. The strength of the correlation (rm-R^2) was determined with the following criteria: poor (0 to 0.49), moderate (0.50 to 0.75), and strong (> 0.75) (Portney, 2020). Correlation plots generated to examine the linear relationship between each model's predictions and the ground truth.

The Intraclass Correlation Coefficient (ICC) was calculated to assess both intra-session (within the same session) and inter-session (between different sessions) reliability. Reliability was determined using the following criteria: < 0.5 (poor), 0.5 – 0.75 (moderate), 0.75 – 0.9 (good), and > 0.9 (excellent) (Koo and Li., 2016). These ICC values provide insights into the consistency of each model's predictions relative to the ground truth. Specifically, inter-session reliability was calculated by comparing the peak reach distances from the five repetitions performed by each participant in session one against the corresponding five repetitions in session two. Additionally, the Coefficient of Variation (COV%) and the Standard Error of Measurement (SEM) were calculated to evaluate the precision and variability of the model outputs. A two-way repeated measures ANOVA was conducted to examine the effects of session and measurement method on reach distance, accounting for individual differences. This test was chosen because the same participants were measured repeatedly across sessions using different methods (v8n, v8x, ground truth). The ANOVA assessed both main effects (session and method) and their interaction while accounting for within-subject variability (Atkinson & Nevill. 1998). Finally, Bland-Altman plots with average mean difference (MD) and 95% limits of agreement (LOA) were used to evaluate agreement between 2D and 3D measures of reaching distance. P-values < 0.05 were considered significant.

3 Results

3.1 Forward Reaching

To ensure the most suitable YOLOv8 model was selected, both the YOLOv8n and YOLOv8x models were evaluated. During forward-reaching sit-and-reach trials, results demonstrated that the YOLOv8n model had a stronger correlation to the ground truth data

(Figure 4). Examining all reach distances using repeated measures correlation revealed a very high correlation for the YOLOv8n model (rm-R^2 = 0.98, $p < .001$) compared to the ground truth, while the YOLOv8x model demonstrated a strong correlation (rm-R^2 = 0.85, $p < .001$). The intra-session reliability of the YOLOv8n was excellent (ICC = 0.98 in Session 1; 0.99 in Session 2), while YOLOv8x's intra-session reliability was good (ICC = 0.86 in Session 1; 0.89 in Session 2) (Table 1). Additionally, the YOLOv8n demonstrated a substantially lower intra-session standard error of measurement (SEM) (24.51 mm and 21.25 mm) when compared to the YOLOv8x (47.94 mm and 46.48 mm). However, the COV was higher for the YOLOv8n model (36.06% and 36.10%) compared to the YOLOv8x model (24.82% and 26.63%) in relation to the ground truth data (Table 1). The inter-session reliability of the YOLOv8n was 0.98, while YOLOv8x's inter-session reliability was 0.87 (95% CI [0.83, 0.91]) (Table 2). Across sessions, YOLOv8n demonstrated a higher overall inter-session SEM (16.16 mm) compared to YOLOv8x (11.03 mm) and a higher inter-session COV (35.98% vs 25.55%). The two-way repeated measures ANOVA revealed no significant main effects or interaction, with generalised eta-squared values indicating trivial effect sizes for session ($F = 1.37$, $p = .263$, $\eta^2g = .005$), method ($F = 1.02$, $p = .374$, $\eta^2g = .003$), and the session × method interaction ($F = 2.72$, $p = .085$, $\eta^2g = .001$).

The Bland-Altman plot for the YOLOv8n versus ground truth measurements indicated a slight positive bias (mean difference: 8.77 mm, Figure 5), with limits of agreement ranging from -50.36 mm to 67.91 mm, reflecting moderate variability between the methods. In contrast, the YOLOv8x versus ground truth measurements demonstrated a larger negative bias (mean difference: -12.28 mm, Figure 5) and substantially wider limits of agreement (-167.25 mm to 142.69 mm), highlighting pronounced discrepancies, particularly at higher measurement values. Two-way repeated measures ANOVA examined the effects of session

and measurement method (YOLOv8n, YOLOv8x, ground truth) on forward reach distance. There was no significant main effect of session, no significant main effect of measurement method, and no significant session-method interaction. Effect sizes were negligible for all effects, indicating that the non-significant findings reflect genuinely small effects rather than insufficient statistical power (Atkinson & Nevill, 1998).

****Figure 4 Here****

Figure 4 shows a correlation between the YOLOv8n (left) and YOLOv8x(right) forward sit-and-reach distances compared to the ground truth data.

****Figure 5 Here****

Figure 5 Bland-Altman plots comparing reach distance measurements from the YOLOv8n (left) and YOLOv8x (right) models with those from a motion capture system during forward-reaching sit-and-reach trials.

3.2 Lateral Reaching

Upon comparison of the lateral-reaching data, results contrasted with the forward-reaching trials; the YOLOv8x model was better correlated to the ground truth data. Taking these findings into account, because each model performed better in one specific task, both models were carried forward for full analysis. When examining all reach distances during lateral reaching sit-and-reach trials, the findings revealed a moderate correlation for both the YOLOv8n and YOLOv8x (Figure 6) models and the ground truth. The intra-session reliability of

the YOLOv8n was poor (ICC = 0.27 in Session 1; 0.22 in Session 2), whilst YOLOv8x's intra-session reliability was slightly better but still poor (ICC = 0.35 in Session 1; 0.27 in Session 2) (Table 1). Furthermore, the inter-session reliability for the YOLOv8x was 0.35, which is a marginal improvement over the YOLOv8n model (ICC = 0.25) (Table 2), though both models experienced a substantial drop compared to forward reaching trials. Contrastingly to forward reaching trials, the YOLOv8n model exhibited a higher overall inter-session SEM (37.23 mm) compared to the YOLOv8x (32.11 mm) during lateral reaching sit-and-reach trials (Table 2). Furthermore, the inter-session coefficient of variation (COV) was again higher for the YOLOv8n model (24.84%) compared to the YOLOv8x model (20.86%).

The Bland-Altman plots for YOLOv8n compared to the ground truth demonstrated a notable negative bias (mean difference: -95.50 mm), with limits of agreement ranging from -159.96 mm to -31.04 mm, indicating considerable variability between the methods (Figure 7). In contrast, the YOLOv8x versus ground truth comparison revealed a smaller negative bias (mean difference: -82.58 mm) and narrower limits of agreement (-135.47 mm to -29.68 mm), suggesting less pronounced discrepancies compared to the YOLOv8n comparison (Figure 7). A two-way repeated measures ANOVA examined the effects of session and measurement method (YOLOv8n, YOLOv8x, ground truth) on lateral reach distance. The two-way repeated measures ANOVA revealed a significant main effect of method ($F = 180.09$, $p < .001$, $\eta^2g = .495$), indicating a large effect whereby reach distance differed substantially between the YOLO models and ground truth; neither the main effect of session ($F = 0.32$, $p = .582$, $\eta^2g = .006$) nor the session × method interaction ($F = 1.36$, $p = .277$, $\eta^2g = .003$) reached significance, with both demonstrating trivial effect sizes. This was further confirmed by the paired t-tests, with both models significantly underestimating lateral reach distance compared to ground

truth. The large effect size indicates that this difference reflects a genuine and substantial systematic bias in the estimation of lateral reach distance, rather than measurement noise.

****Figure 6 Here****

Figure 6 shows a correlation between the YOLOv8n (left) and YOLOv8x(right) for lateral sit-and-reach distances compared to the ground truth data.

****Figure 7 Here****

Figure 7 Bland-Altman plots comparing reach distance measurements from the YOLOv8n (left) and YOLOv8x (right) models with those from a motion capture system during lateral-reaching sit-and-reach trial.

Table 1 Intra-session reliability and descriptive statistics of the YOLOv8n and YOLOv8x model compared to ground truth during sit-and-reach tasks.

	Session 1				***Session 2***			
	Forward Reaching		**Lateral Reaching**		**Forward Reaching**		**Lateral Reaching**	
Model	YOLOv8n	YOLOv8x	YOLOv8n	YOLOv8x	YOLOv8n	YOLOv8x	YOLOv8n	YOLOv8x
Stats								
Mean (mm)	524.82	516.76	174.58	189.4	557.95	522.96	169.92	180.44
SD (mm)	189.27	128.27	44.34	40.19	201.45	139.25	41.16	36.2
ICC	0.98 (0.97 - 0.99)	0.86 (0.79 - 0.91)	0.27 (-0.05 - 0.64)	0.35 (-0.05 - 0.71)	0.99 (0.98 - 0.99)	0.89 (0.81 - 0.93)	0.22 (-0.05 - 0.57)	0.27 (-0.04 - 0.64)
SEM (mm)	24.51	47.94	37.79	32.48	21.25	46.48	36.33	30.94
COV (%)	36.06	24.82	25.4	21.22	36.1	26.63	24.22	20.06

Table 2 Inter-session reliability of the YOLOv8n and YOLOv8x model compared to ground truth during sit-and-reach tasks.

	Forward Reaching		*Lateral Reaching*	
Model	YOLOv8n	YOLOv8x	YOLOv8n	YOLOv8x
Stats				
Mean (mm)	540.83	519.81	172.54	185.25
SD (mm)	195.34	133.28	42.91	38.63
ICC	0.98 (0.98-0.99)	0.87 (0.83 - 0.9)	0.25 (-0.05 - 0.6)	0.35 (-0.05 - 0.68)
SEM (mm)	16.16	11.03	37.23	32.11
COV (%)	35.98	25.55	24.84	20.86

4 Discussion

This study aimed to evaluate the inter- and intra-session reliability and concurrent validity of the YOLOv8x and YOLOv8n HPE models for assessing the functional reach test compared to 3D marker-based motion capture. The findings of this study partially support our hypothesis, revealing that model performance is task dependent. Specifically, during forward-reaching tasks, the YOLOv8n model demonstrated excellent intra- and inter-session reliability alongside strong concurrent validity, whereas the YOLOv8x model exhibited good, though comparatively lower, reliability. Conversely, both models demonstrated poor reliability and wider limits of agreement during lateral-reaching tasks. These task-specific discrepancies highlight crucial nuances regarding model complexity, measurement error, and the impact of physical occlusions, which warrant further exploration for future clinical applications.

4.1 Forward Reaching

The results demonstrated that YOLOv8n and YOLOv8x models have strong correlation to the ground truth data in the forward reaching trials, which improved for both models when the 3 largest reach distances for each participant were correlated with the ground truth. The results demonstrated excellent inter- and intra-session reliability for the YOLOv8n model, whereas the YOLOv8x model demonstrated good reliability during forward reaching sit-and-reach trials. The YOLOv8n and YOLOv8x models also demonstrate notable differences in the SEM, YOLOv8n shows a substantially lower intra-session SEM compared to YOLOv8x, indicating that the lighter model (V8n) has better within-session precision. Furthermore, YOLOv8n demonstrated an improvement in accuracy over time, with the intra-session SEM decreasing from Session 1 to Session 2, highlighting its sustained capability across repeated

trials. When examining the COV, YOLOv8x did show a relatively stable profile across time, likely because the increased depth of the model allows for some generalisability across sessions (Terven et al., 2023). However, this specific COV consistency must be interpreted cautiously, particularly given the larger magnitude of measurement error observed in other metrics for this model. Bland-Altman plots raise important considerations of the clinical utility of YOLOv8n and YOLOv8x in relation to forward-reach sit-and-reach measurements. YOLOV8n showed relatively less bias and narrower limits of agreement, reflecting a strong potential for providing measurements which are close to the true value across repeated trials. However, the larger variability, wider limits of agreement, and negative bias observed with YOLOv8x raise concerns regarding its accuracy in capturing forward-reaching movements, especially when measurements extend beyond average ranges.

When interpreting these findings, it is essential to draw a clear distinction between correlation and agreement. Although YOLOv8n demonstrated excellent reliability, the Bland-Altman results illustrate that it does not always yield exact millimetre precision. While YOLOv8n demonstrates strong concurrent validity in tracking relative performance changes, its absolute spatial agreement is still subject to bias. Although the mean difference between the YOLOv8n/x and 3D measurements was below the minimal detectable change range of 3.7–5.2 cm, this threshold is based on clinical populations with balance or mobility impairments (Lynch et al., 1998; Katz-Leurer et al., 2009). Finally, as this study involved healthy young adults, this comparison should be interpreted with caution. Nevertheless, it offers a useful benchmark to gauge the potential for future clinical applications.

Sit-and-reach tests are well-established methods to assess dynamic sitting balance and postural control, as shown by the relationship between seated forward reach distance and

clinical balance measures in spinal cord injury and stroke (Field-Fote & Ray, 2010; Chiou & Strutton, 2020; van Helden et al., 2023). Recent research in spinal rehabilitation has employed markerless motion analysis to assess movements of the trunk and limbs for example, trunk flexion during sit-and-reach tasks with a high degree of accuracy (Sohan et al., 2024). Markerless motion analysis systems are often favoured over traditional motion-capture for rehabilitation due to their ease of setup, affordability, and ability to be used outside the lab (Sliwinski et al., 2020; Joshi & Ganvir, 2022). Coupled with its overall speed and scalability, the strong concurrent validity and high reliability of YOLOv8n in forward-reaching demonstrate its potential effectiveness as a portable solution for enhancing rehabilitation assessments (Katz-Leurer et al., 2009; Terven et al., 2023).

YOLOv8 models were used to measure reaching distance of the upper limbs in this study, the model provides estimates for all 17 body joints, including the trunk and lower limbs. This means the same model could potentially be used to quantify movement across multiple body segments during reaching or other functional tasks. We focused on reaching distance because it is a widely used clinical metric of balance and upper limb function; however, YOLOv8's full-body tracking capability may allow clinicians to assess changes in motor control strategies, such as increased trunk inclination, that could influence reaching performance in clinical populations. While this application was beyond the scope of the current study, future work should evaluate the reliability and validity of additional kinematic metrics from YOLOv8 in individuals with motor impairments. Overall, the outcomes of the study show that YOLOv8n demonstrates greater consistency and reliability during forward-reaching sit-and-reach tasks, supported by higher ICC values and a lower intra-session SEM compared to YOLOv8x. While the strong performance of YOLOv8n suggests it may be beneficial for tracking physical

performance over time, its specific applicability and reliability for true longitudinal studies must be directly evaluated in future research.

4.2 Lateral Reaching

The results demonstrated that YOLOv8n and YOLOv8x models showed only a moderate correlation to the ground truth data during lateral reaching sit-and-reach trials. Test-retest reliability for lateral reaching trials is notably lower compared to forward reaching trials, with both intra-session and inter-session ICC values for both models falling within the poor range. The YOLOv8n model shows a higher SEM compared to the YOLOv8x across both intra-session and inter-session measurements, indicating that YOLOV8n experiences decreased accuracy when estimating lateral reach distances. This trend is consistent across sessions, suggesting that the less complex architecture of YOLOv8n fares worse in measuring lateral movements compared to the larger YOLOv8x model. Both models show similar variability as noted by the CoV, but YOLOv8x tends to post slightly lower COV across sessions. Despite the overall reduced accuracy and poor test-retest reliability observed for both models in lateral trials, the higher complexity of the YOLOv8x model allows it to maintain slightly better relative stability (Terven et al., 2023).

Bland-Altman analysis shows a negative bias, suggesting both models typically underestimate reaching distances in lateral sit-and-reach trials compared to the ground truth. Although YOLOv8x demonstrates less variability and a comparatively smaller bias than YOLOv8n, the limits of agreement remain relatively wide for both models. Both models are currently restricted in their capacity to accurately reflect lateral-reaching movements and neither model presently demonstrates the necessary reliability for this specific task. However, the comparative advantage of YOLOv8x does provide insight for future development. The

deeper and more sophisticated architecture of the YOLOv8x model is particularly equipped to handle occlusions, which were common during lateral data collection due to the presence of the chair (Terven et al., 2023). Further refinement of these models is required to improve lateral reaching reliability.

While the YOLOv8x model demonstrated slightly better performance during the chair-based modified sit-and-reach test because its deeper neural network architecture enhances feature extraction and helps interpret partially obscured joint locations, it still struggles with lateral tasks (Dong and Du., 2024). Consequently, the assessment of lateral-reaching movements currently lacks the necessary reliability for clinical application with either model. Regardless, these models demonstrate the potential to eventually enrich clinical assessment by providing objective measures of dynamic sitting balance. With further development to address the current limitations in lateral movements, deploying models like YOLOv8 for dynamic sit-and-reach assessments could allow clinicians to objectively measure reach distances and trunk stability over time. This would support data-driven interventions and potential longitudinal tracking of patient progress in a convenient, non-invasive manner. Further refinement of these models is required to improve lateral reaching reliability.

5 Limitations

A methodological limitation of this study is the lack of strict temporal synchronisation between the 2D and 3D systems, starting frames were identified independently, the peak reach distances compared between the two systems may be slightly offset temporally. This mismatch could artificially inflate our measurement error and the SEM and limits of agreement may be estimates of the models' true spatial agreement. Another limitation of our

video-based HPE analysis is its dependency on recording frame rate. 25 Hz sampling rate of the 2D video offered significantly lower temporal resolution compared to the 250 Hz 3D optical motion capture system. Consequently, when participants were asked to pause to facilitate the automatic identification of movement onset the 25 Hz rate caused temporal smoothing of the kinematics. This combined with inconsistent pausing by participants resulted in the manual selection of starting frames. This may introduce operator dependency partially negating the fundamental advantage of a fully automated markerless assessment outlined in previous research (Marchesi et al., 2021; Dotti et al., 2024; Ceriello et al., 2025).

To improve temporal precision and eliminate operator dependency, we strongly recommend future studies utilize higher frame rates. Consumer-grade devices, such as smartphone cameras operating at 60 Hz and are practical and cost effective for clinic or home-based environments. Methodologically, enforcing strict, prolonged stationary postures prior to movement will enable automated scripts to reliably apply velocity-based segmentation without manual intervention. Furthermore, integrating supplementary computer vision techniques, such as background subtraction, alongside YOLOv8 models could enhance tracking fidelity and test-retest reliability when tracking fast movements in cluttered environments (Chien et al., 2024). Finally, because this study was conducted solely on healthy individuals to establish a controlled baseline, the findings lack immediate generalizability. Clinical populations frequently present with altered movement patterns or spasticity; therefore, future studies must validate this tool across diverse patient groups before widespread integration into rehabilitation practice.

6 Conclusion

Our findings suggest that the performance of YOLOv8-based HPE models is highly task-dependent when assessing functional reach in healthy young adults. Both models exhibited strong correlation with ground truth data during forward-reaching tasks, with YOLOv8n demonstrating superior test-retest reliability and lower intra-session measurement error. Conversely, correlation was only moderate during lateral-reaching tasks, where both models exhibited poor test-retest reliability and wide limits of agreement. While the more complex architecture of YOLOv8x appeared to offer slightly better resilience to occluded movements during lateral reaching, neither model currently demonstrates adequate reliability for clinical assessment in this specific plane of motion. Given that our results are based on a sample of healthy participants, further research involving clinical populations, alongside continued algorithmic refinement to improve lateral tracking, is necessary to confirm and expand upon these preliminary observations.

Acknowledgements. We thank Dr Joeri Van Helden, Humain Choudhury, Kamili Bell, and Oluwaseye Olusanya for their assistance with data collection. We are also grateful to all the participants for their time and contribution to the study

Funding: This research received no external funding.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available on request from the corresponding author.

Declaration of Competing Interests: The authors declare no conflict of interest.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the University of Birmingham Research Ethics Committee (ERN_20-1453).

Authorship Contribution Statement: Cavan Aulton was responsible for the primary conceptualisation, methodology, software development, validation, formal analysis, investigation, and writing the original draft. Dr. Chuang-Yuan Chiu contributed to the conceptualization, methodology, software, and supervised the project. Dr. Shin-Yi Chiou contributed to the conceptualization and provided supervision. All authors participated in the review and editing of the final manuscript.

7 References

Alruwaili, M., Atta, M. N., Siddiqi, M. H., Khan, A., Khan, A., Alhwaiti, Y., & Alanazi, S. (2024). Deep learning-based YOLO models for the detection of people with disabilities. IEEE Access, 12, 2543–2566. https://doi.org/10.1109/ACCESS.2023.3347169

Atkinson, G., & Nevill, A. M. (1998). Statistical methods for assessing measurement error (reliability) in variables relevant to sports medicine. Sports Medicine, 26(4), 217–238. https://doi.org/10.2165/00007256-199826040-00002

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. Frontiers in Psychology, 8, Article 456. https://doi.org/10.3389/fpsyg.2017.00456

Baltaci, G., Un, N., Tunay, V., Besler, A., & Gerçeker, S. (2003). Comparison of three different sit and reach tests for measurement of hamstring flexibility in female university students. British Journal of Sports Medicine, 37(1), 59–61. https://doi.org/10.1136/bjsm.37.1.59

Boualia, S. N., & Essoukri Ben Amara, N. (2019). Pose-based human activity recognition: A review. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC) (pp. 1468–1475). https://doi.org/10.1109/IWCMC.2019.8766694

Ceriello, I., Ghislieri, M., Rum, L., Camomilla, V., Macaluso, A., & Borzuola, R. (2026). Muscle coordination strategies during functional reach across multiple directions in healthy individuals. European Journal of Applied Physiology, 126(2), 693–711. https://doi.org/10.1007/s00421-025-05951-7

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. arXiv. https://doi.org/10.48550/arXiv.1908.10357

Chien, C.-T., Ju, R.-Y., Chou, K.-Y., & Chiang, J.-S. (2024). YOLOv8-AM: YOLOv8 with attention mechanisms for pediatric wrist fracture detection. arXiv. https://doi.org/10.48550/arXiv.2402.09329

Chiou, S.-Y., & Strutton, P. H. (2020). Crossed corticospinal facilitation between arm and trunk muscles correlates with trunk control after spinal cord injury. Frontiers in Human Neuroscience, 14, Article 583579. https://doi.org/10.3389/fnhum.2020.583579

Colyer, S. L., Evans, M., Cosker, D. P., & Salo, A. I. T. (2018). A review of the evolution of vision-based motion analysis and the integration of advanced computer vision methods towards developing a markerless system. Sports Medicine - Open, 4, Article 24. https://doi.org/10.1186/s40798-018-0139-y

Crenna, F., Rossi, G. B., & Berardengo, M. (2021). Filtering biomechanical signals in movement analysis. Sensors, 21(13), Article 4580. https://doi.org/10.3390/s21134580

Dong, C., & Du, G. (2024). An enhanced real-time human pose estimation method based on modified YOLOv8 framework. Scientific Reports, 14, Article 8012. https://doi.org/10.1038/s41598-024-58146-z

Dotti, G., Caruso, M., Fortunato, D., Knaflitz, M., Cereatti, A., & Ghislieri, M. (2024). A statistical approach for functional reach-to-grasp segmentation using a single inertial measurement unit. Sensors, 24(18), Article 6119. https://doi.org/10.3390/s24186119

Fang, H.-S., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., Li, Y.-L., & Lu, C. (2023). AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(6), 7157–7173. https://doi.org/10.1109/TPAMI.2022.3222784

Field-Fote, E. C., & Ray, S. S. (2010). Seated reach distance and trunk excursion accurately reflect dynamic postural control in individuals with motor-incomplete spinal cord injury. Spinal Cord, 48(10), 745–749. https://doi.org/10.1038/sc.2010.11

Francia, C., Donno, L., Motta, F., Cimolin, V., Galli, M., & LoMauro, A. (2026). Accuracy and Reliability of Markerless Human Pose Estimation for Upper Limb Kinematic Analysis Across Full and Partial Range of Motion Tasks. *Applied Sciences*, *16*(3). https://doi.org/10.3390/app16031202

Garrido-Castro, J. L., Medina-Carnicer, R., Schiottis, R., Galisteo, A. M., Collantes-Estevez, E., & Gonzalez-Navas, C. (2012). Assessment of spinal mobility in ankylosing spondylitis using a video-based motion capture system. Manual Therapy, 17(5), 422–426. https://doi.org/10.1016/j.math.2012.03.011

Hansen, R. M., Arena, S. L., & Queen, R. M. (2024). Validation of upper extremity kinematics using markerless motion capture. Biomedical Engineering Advances, 7, Article 100128. https://doi.org/10.1016/j.bea.2024.100128

Hashemi, S. M. H., Safari, L., & Taromi, A. D. (2024). Realism in action: Anomaly-aware diagnosis of brain tumors from medical images using YOLOv8 and DeiT. arXiv. https://doi.org/10.48550/arXiv.2401.03302

Henrichsen, A. (2000). 3D reconstruction and camera calibration from 2D images (Technical Report). Technical University of Denmark

Hornsby, E. A., & Johnston, L. M. (2021). Evaluating the impact of a Pilates intervention on physical function in children with hypermobility spectrum disorder: A study protocol using single-case experimental design. Open Journal of Pediatrics, 11(1), 55–70. https://doi.org/10.4236/ojped.2021.111006

Jakob, V., Küderle, A., Kluge, F., Klucken, J., Eskofier, B. M., Winkler, J., Winterholler, M., & Gassner, H. (2021). Validation of a sensor-based gait analysis system with a gold-standard motion capture system in patients with Parkinson's disease. Sensors, 21(22), Article 7680. https://doi.org/10.3390/s21227680

Joshi, K., & Ganvir, S. (2022). Correlation between modified functional reach test and transfer activities of daily living in individuals with spinal cord injury - a pilot study. Journal of Medical Research, 8(4), 135–138. https://doi.org/10.31254/jmr.2022.8404

Katz-Leurer, M., Fisher, I., Neeb, M., Schwartz, I., & Carmeli, E. (2009). Reliability and validity of the modified functional reach test at the sub-acute stage post-stroke. Disability and Rehabilitation, 31(3), 243–248. https://doi.org/10.1080/09638280801927830

Khalaf, A. T., & Abdulateef, S. K. (2024). Ophthalmic diseases classification based on YOLOv8. Journal of Robotics and Control, 5(2). https://doi.org/10.18196/jrc.v5i2.21208

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine, 15(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lee, Y.-C., Yu, X., & Xiong, W. (2022). A comparative evaluation of the four measurement methods for comfort and reachability distance perceptions. Behavior Research Methods, 54(4), 1766–1777. https://doi.org/10.3758/s13428-021-01715-1

Leporace, G., Metsavaht, L., Gonzalez, F. F., Arcanjo de Jesus, F., Machado, M., Celina Guadagnin, E., & Gomes-Neto, M. (2023). Validity and reliability of two-dimensional video-based assessment to measure joint angles during running: A systematic review and meta-analysis. Journal of Biomechanics, 157, Article 111747. https://doi.org/10.1016/j.jbiomech.2023.111747

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Computer Vision – ECCV 2014 (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Lynch, S. M., Leahy, P., & Barker, S. P. (1998). Reliability of measurements obtained with a modified functional reach test in subjects with spinal cord injury. Physical Therapy, 78(2), 128–133. https://doi.org/10.1093/ptj/78.2.128

Marchesi, G., Ballardini, G., Barone, L., Giannoni, P., Lentino, C., De Luca, A., & Casadio, M. (2021). Modified functional reach test: Upper-body kinematics and muscular activity in chronic stroke survivors. Sensors, 22(1), Article 230. https://doi.org/10.3390/s22010230

Rana, M. S. (2024). Leveraging markerless computer vision for comprehensive walking automated gait analysis in rehabilitation [Bachelor's thesis, Arcada University of Applied Sciences]. Theseus. http://www.theseus.fi/handle/10024/860826

Reinking, M. F., Dugan, L., Ripple, N., Schleper, K., Scholz, H., Spadino, J., Stahl, C., & McPoil, T. G. (2018). Reliability of two-dimensional video-based running gait analysis. International Journal of Sports Physical Therapy, 13(3), 453–461.

Scataglini, S., Abts, E., Van Bocxlaer, C., Van den Bussche, M., Meletani, S., & Truijen, S. (2024). Accuracy, validity, and reliability of markerless camera-based 3D motion capture systems versus marker-based 3D motion capture systems in gait analysis: A systematic review and meta-analysis. Sensors, 24(11), Article 3686. https://doi.org/10.3390/s24113686

Schenkman, M., Cutson, T. M., Kuchibhatla, M., Chandler, J., & Pieper, C. (1997). Reliability of impairment and physical performance measures for persons with Parkinson's disease. Physical Therapy, 77(1), 19–27. https://doi.org/10.1093/ptj/77.1.19

Schurr, S. A., Marshall, A. N., Resch, J. E., & Saliba, S. A. (2017). Two-dimensional video analysis is comparable to 3D motion capture in lower extremity movement assessment. International Journal of Sports Physical Therapy, 12(2), 163–172.

Sliwinski, M. M., Akselrad, G., Alla, V., Buan, V., & Kaemmerlen, E. (2020). Community exercise programing and its potential influence on quality of life and functional reach for individuals with spinal cord injury. The Journal of Spinal Cord Medicine, 43(3), 358–363. https://doi.org/10.1080/10790268.2018.1543104

Sohan, M., Sai Ram, T., & Rami Reddy, C. V. (2024). A review on YOLOv8 and its advancements. In Data Intelligence and Cognitive Informatics (pp. 529–545). Springer. https://doi.org/10.1007/978-981-99-7962-2_39

Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS. Machine Learning and Knowledge Extraction, 5(4). https://doi.org/10.3390/make5040083

van Helden, J. F. L., Alexander, E., Cabral, H. V., Strutton, P. H., Martinez-Valdes, E., Falla, D., Chowdhury, J. R., & Chiou, S.-Y. (2023). Home-based arm cycling exercise improves trunk control in persons with incomplete spinal cord injury: An observational study. Scientific Reports, 13, Article 22120. https://doi.org/10.1038/s41598-023-49053-w

Wallmann, J. (2023). Classifying physical exercises and counting repetitions using three-dimensional pose estimation [Bachelor's thesis, University of Applied Sciences Technikum Wien].

Widhalm, K., Durstberger, S., Greisberger, A., Wolf, B., & Putz, P. (2024). Validity of assessing level walking with the 2D motion analysis software TEMPLO and reliability of 3D marker application. Scientific Reports, 14, Article 1427. https://doi.org/10.1038/s41598-024-52053-z

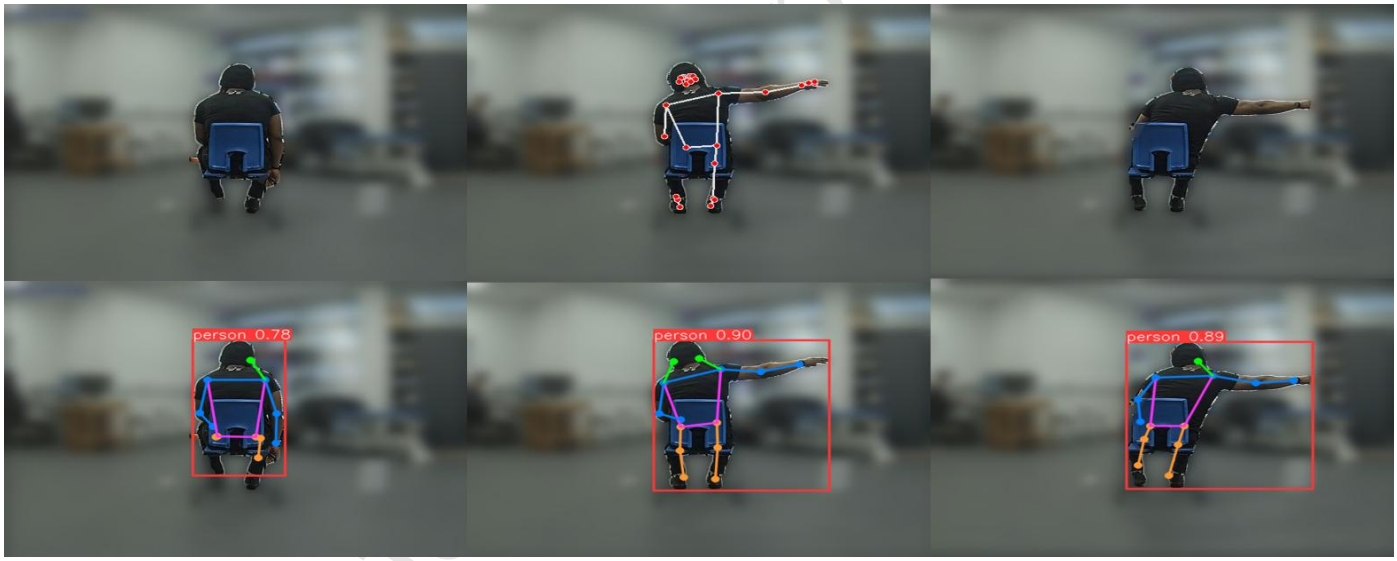
Zhang, J., Zhang, D., Xu, X., Jia, F., Liu, Y., Liu, X., Ren, J., & Zhang, Y. (2020). MobiPose: Real-time multi-person pose estimation on mobile devices. In Proceedings of the 18th Conference on

Embedded Networked Sensor Systems (pp. 136–149). Association for Computing Machinery. https://doi.org/10.1145/3384419.3430726

Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., & Shah, M. (2022). Deep learning-based human pose estimation: A survey. arXiv. https://doi.org/10.48550/arXiv.2012.13392

8 Appendix

8.1 Appendix 1



Appendix Figure 1. Visual comparison of three key frames from a lateral reaching trial, analysed by both models. The top row displays frames processed by the MediaPipe model, while the bottom row shows frames analysed by the YOLOv8 model.

Note: The MediaPipe model is unable to detect the person in 2 out of 3 of the key frames and this was common through the dataset, the background has been blurred for presentational reasons. These findings gave us strong evidence to choose a different model for the HPE of this study.

8.2 Appendix 2

Appendix Table 1. Statistical comparison between manually detected starting frames and the velocity-based detection approach described for forward and lateral reaching trials in the Methods section for HPE data.

Metric	Forward Reaching	Lateral Reaching
Mean diff (auto − manual)	+1.99 frames	−5.61 frames
Mean absolute diff	9.76 frames	15.64 frames
Median absolute diff	8 frames	10 frames
Std	12.71 frames	23.12 frames
Max absolute diff	43 frames	141 frames

8.3 Appendix 3

The primary researcher independently identified the movement onset frame for each trial on two separate occasions, with a 1-day interval between sessions to minimise recall bias. Agreement between the two rating sessions was evaluated using a two-way mixed-effects, absolute agreement, single-rater intraclass correlation coefficient (ICC(3,1)), as recommended by Koo and Li (2016) for assessing the consistency of a single rater across repeated

measurements. ICC(3,1) was selected over consistency-based models as absolute agreement was required; that is, the researcher must have identified the same frame, not merely ranked trials in the same relative order. Analysis was conducted on a sample of 10 trials (equalling 49 individual peaks as one randomly selected trial contained 4 peaks) . Results demonstrated excellent intra-rater reliability (ICC(3,1) = 0.990, 95% CI [0.98, 0.99], $p < .001$), indicating that the manual identification of movement onset frames was highly reproducible and that the findings reported in this study are not attributable to rater inconsistency.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Figure 1

Figure 2

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

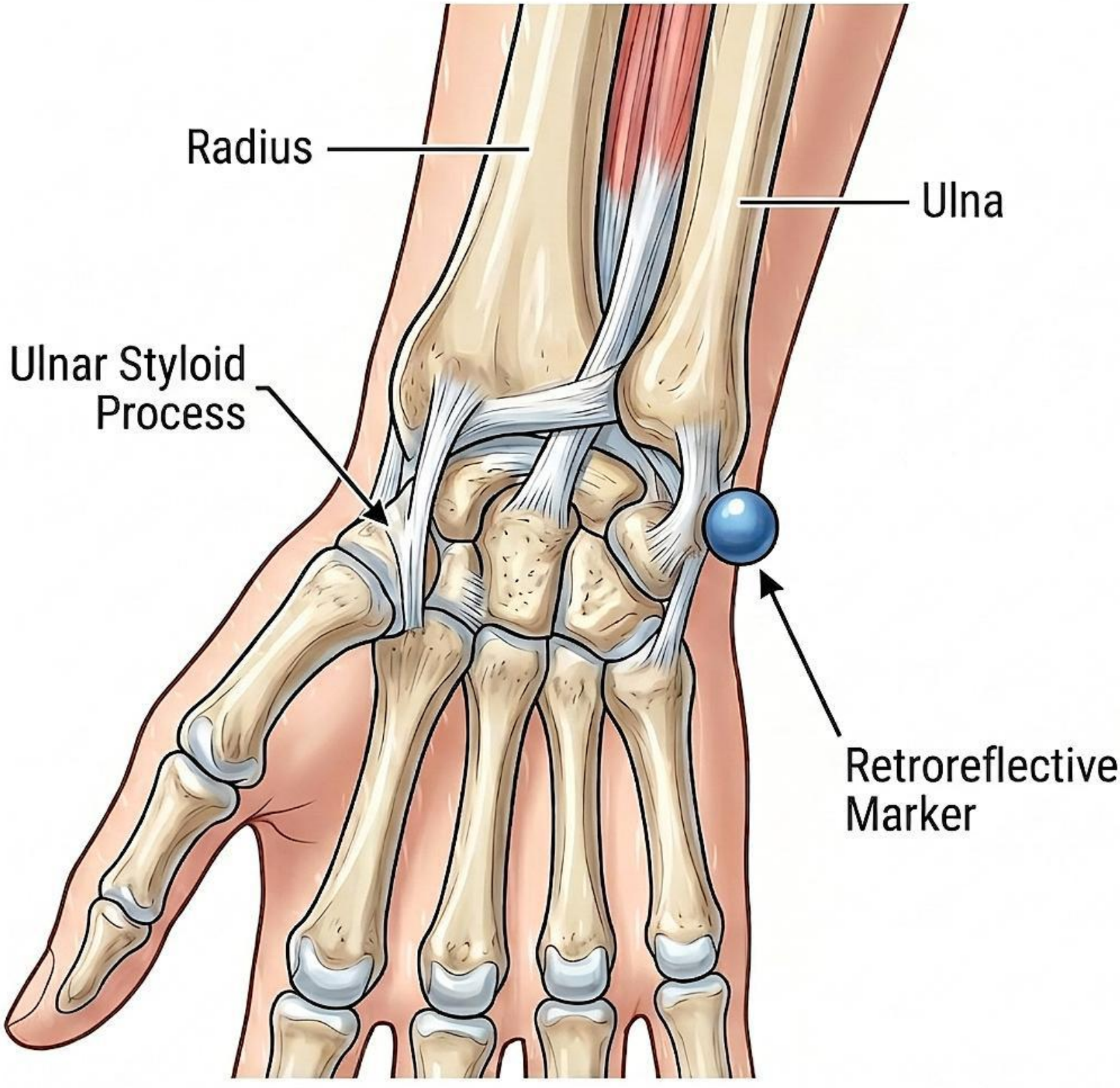

Radius
Ulna
Ulnar Styloid Process
Retroreflective Marker

Figure 1

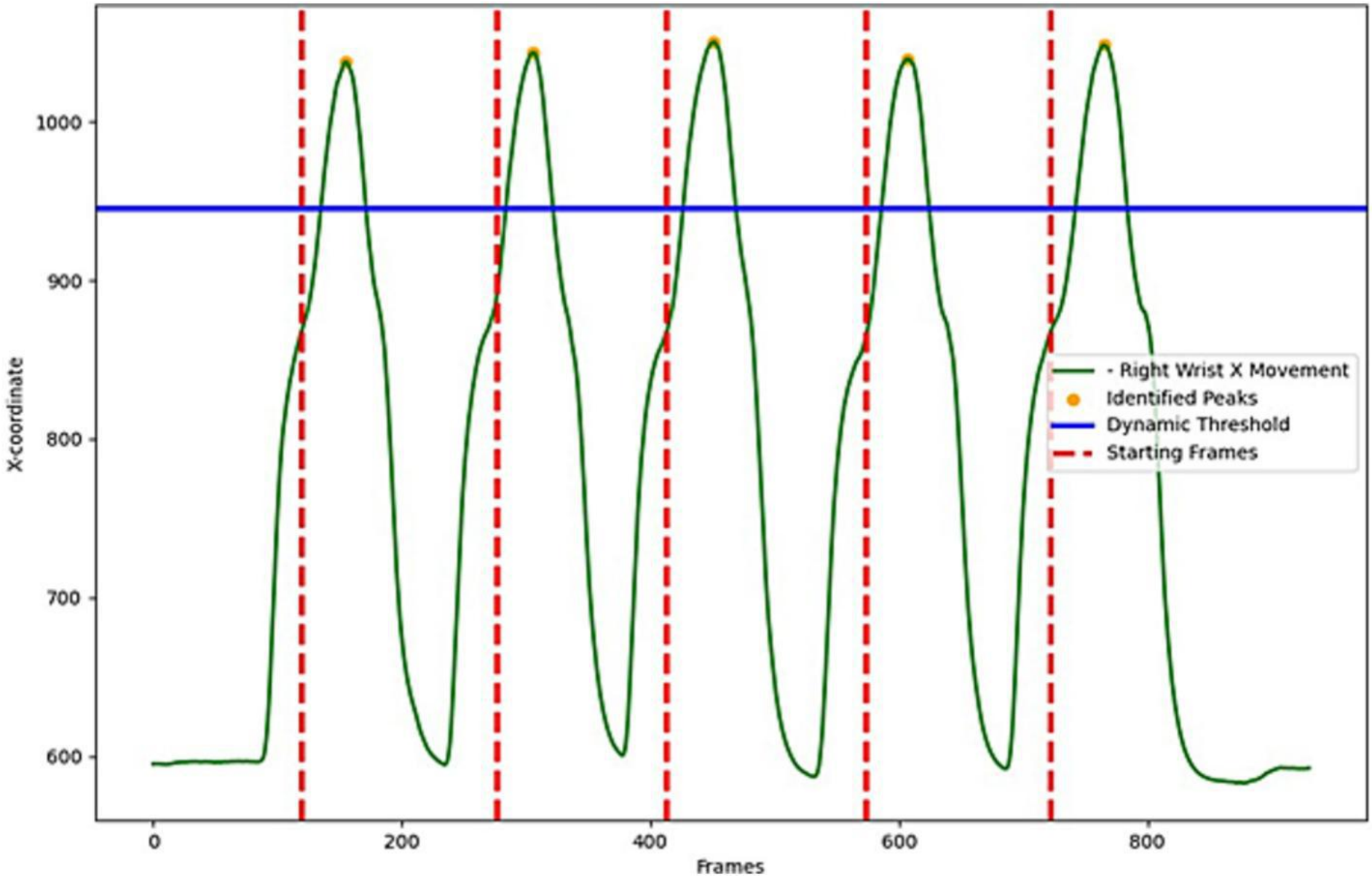

1000
900
800
700
600
X-coordinate
0
200
400
600
800
Frames
- Right Wrist X Movement
Identified Peaks
Dynamic Threshold
Starting Frames

Figure 2

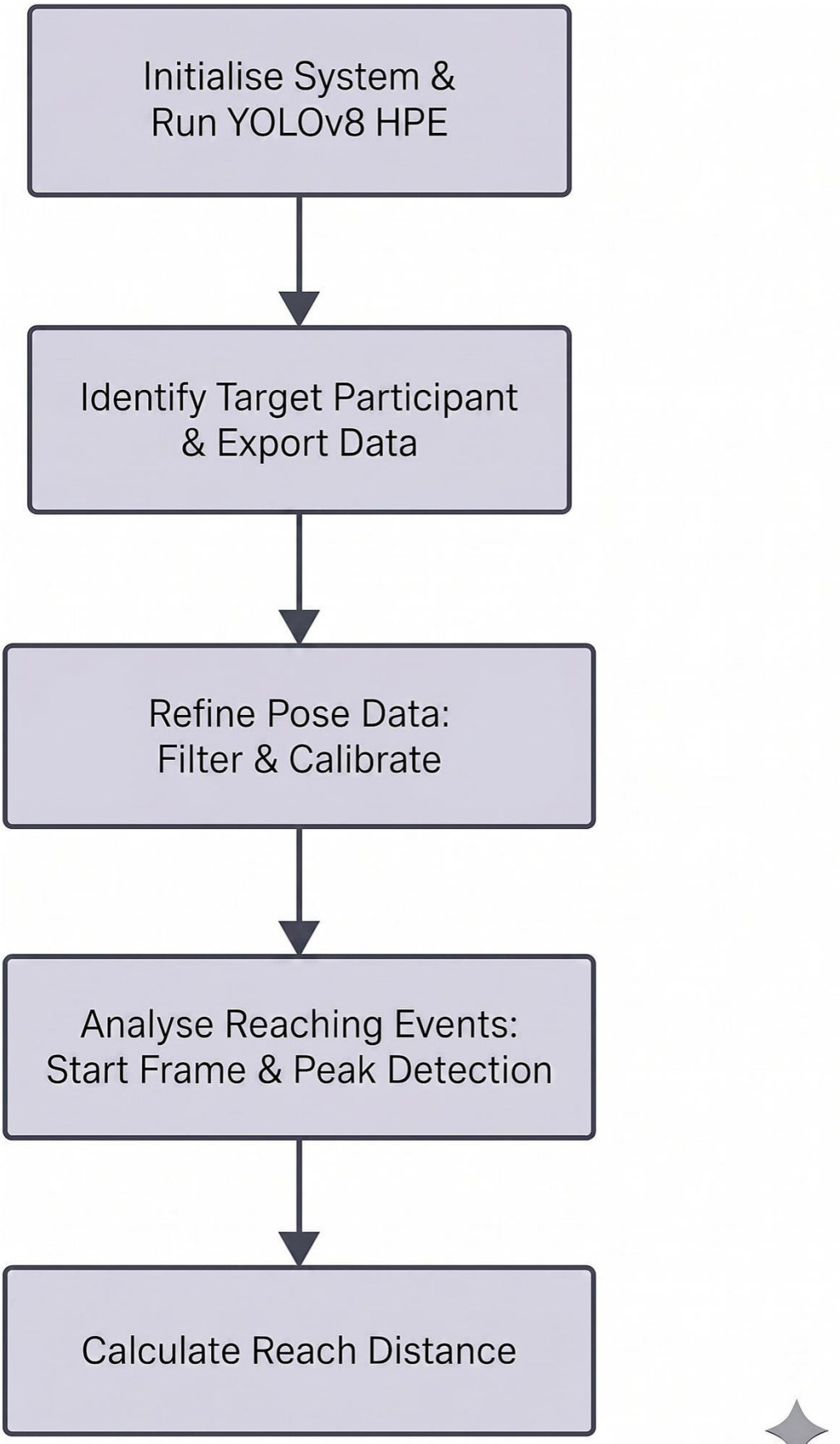
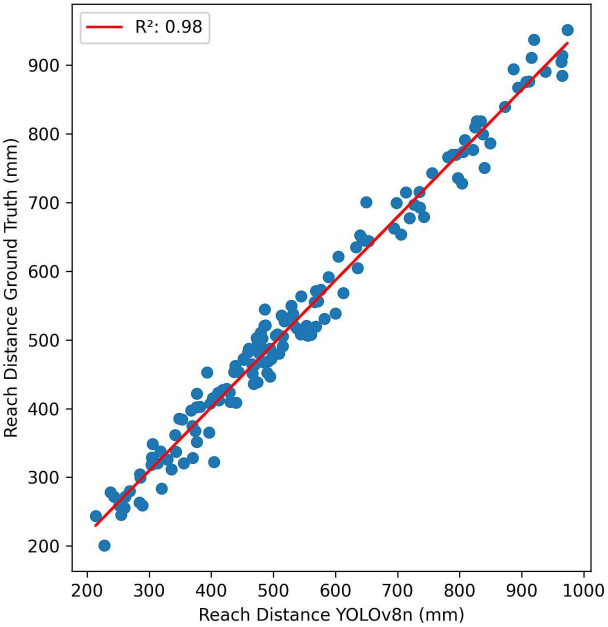

Initialise System &
Run YOLOv8 HPE
Identify Target Participant
& Export Data
Refine Pose Data:
Filter & Calibrate
Analyse Reaching Events:
Start Frame & Peak Detection
Calculate Reach Distance

Figure 3


R²: 0.98
Reach Distance Ground Truth (mm)
Reach Distance YOLOv8n (mm)
200
300
400
500
600
700
800
900
1000

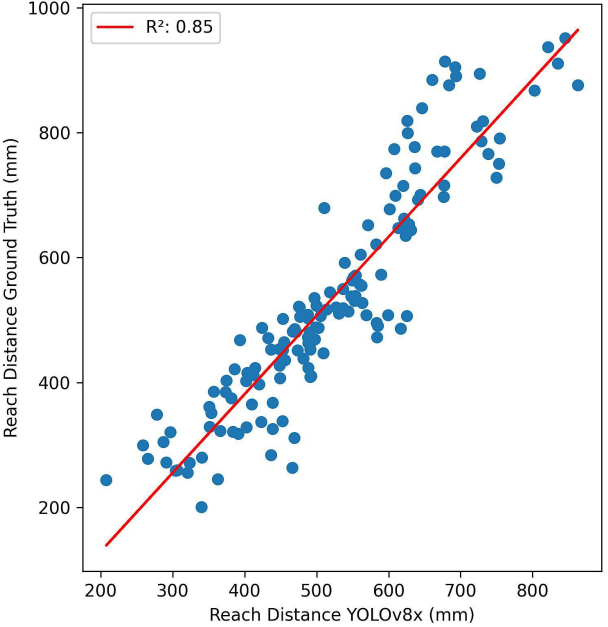
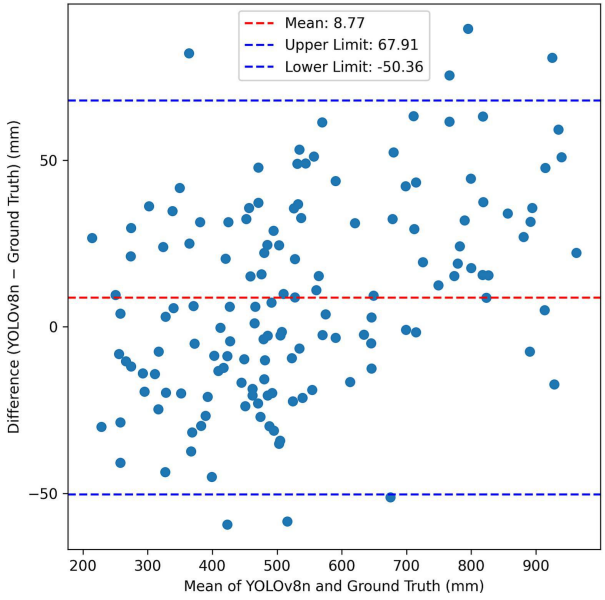

R²: 0.85
Reach Distance Ground Truth (mm)
Reach Distance YOLOv8x (mm)
200
400
600
800
1000

Figure 4


Mean: 8.77
Upper Limit: 67.91
Lower Limit: -50.36
Difference (YOLOv8n − Ground Truth) (mm)
Mean of YOLOv8n and Ground Truth (mm)

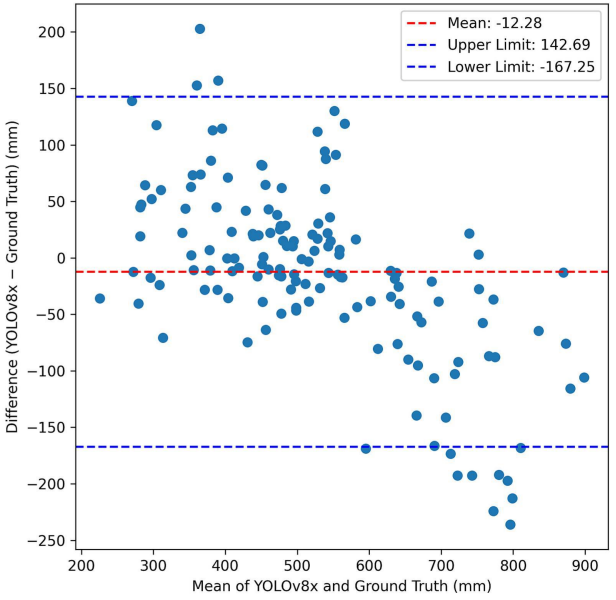
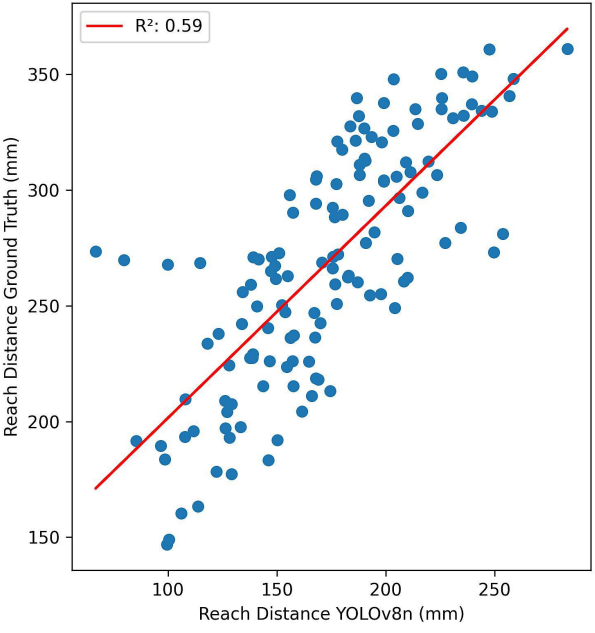

Mean: -12.28
Upper Limit: 142.69
Lower Limit: -167.25
Difference (YOLOv8x − Ground Truth) (mm)
Mean of YOLOv8x and Ground Truth (mm)

Figure 5


R²: 0.59
Reach Distance Ground Truth (mm)
Reach Distance YOLOv8n (mm)
150
200
250
300
350
100
150
200
250

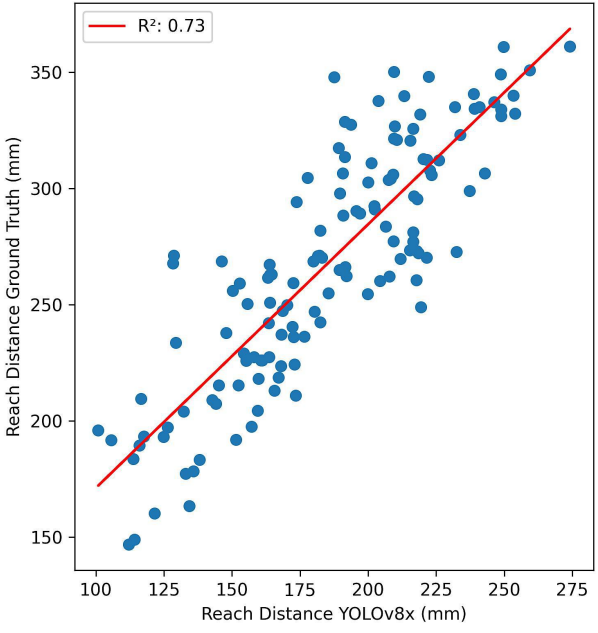
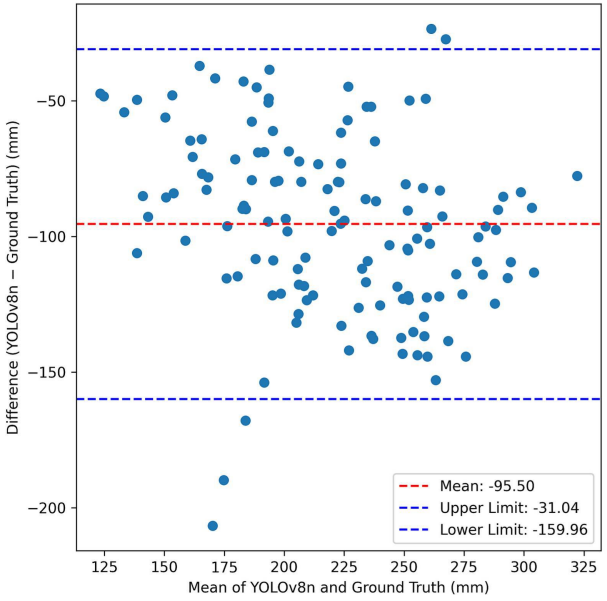

R²: 0.73
Reach Distance Ground Truth (mm)
Reach Distance YOLOv8x (mm)
150
200
250
300
350
100
125
150
175
200
225
250
275

Figure 6


Difference (YOLOv8n − Ground Truth) (mm)
−50
−100
−150
−200
Mean: -95.50
Upper Limit: -31.04
Lower Limit: -159.96
125
150
175
200
225
250
275
300
325
Mean of YOLOv8n and Ground Truth (mm)

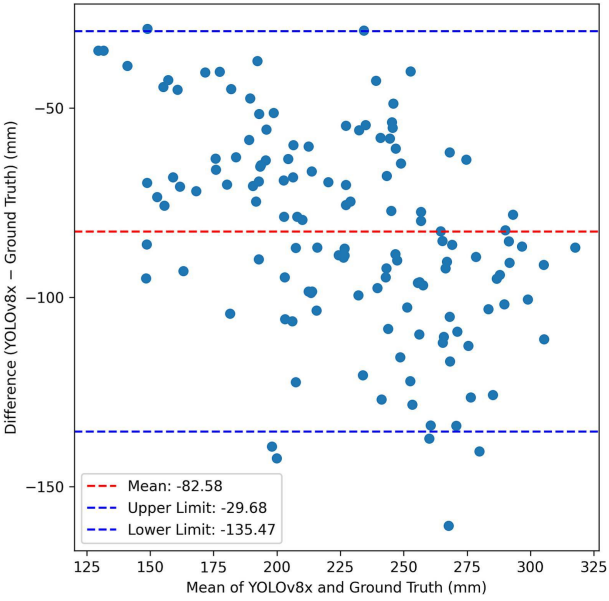

Difference (YOLOv8x − Ground Truth) (mm)
−50
−100
−150
Mean: -82.58
Upper Limit: -29.68
Lower Limit: -135.47
125
150
175
200
225
250
275
300
325
Mean of YOLOv8x and Ground Truth (mm)

Figure 7