

# Explaining rifle shooting factors through multi-sensor body tracking<sup>1</sup>

*Using transformers and attention to mine actionable patterns from skeleton graphs*

Jonatan Flyckt<sup>a,\*</sup>, Filip Andersson<sup>a</sup>, Florian Westphal<sup>a</sup>, Andreas Månsson<sup>b</sup> and Niklas Lavesson<sup>c</sup>

<sup>a</sup>*Department of Computing, Jönköping University, Jönköping, Sweden*

<sup>b</sup>*Saab AB, Training and Simulation, Huskvarna, Sweden*

<sup>c</sup>*Department of Software Engineering, Blekinge Institute of Technology, Sweden*

**Abstract.** There is a lack of data-driven training instructions for sports shooters, as instruction has commonly been based on subjective assessments. Many studies have correlated body posture and balance to shooting performance in rifle shooting tasks, but have mostly focused on single aspects of postural control. This study has focused on finding relevant rifle shooting factors by examining the entire body over sequences of time. A data collection was performed with 13 human participants carrying out live rifle shooting scenarios while being recorded with multiple body tracking sensors. A pre-processing pipeline produced a novel skeleton sequence representation, which was used to train a transformer model. The predictions from this model could be explained on a per sample basis using the attention mechanism, and visualised in an interactive format for humans to interpret. It was possible to separate the different phases of a shooting scenario from body posture with a high classification accuracy (80%). Shooting performance could be detected to an extent by separating participants using their strong and weak shooting hand. The dataset and pre-processing pipeline, as well as the techniques for generating explainable predictions presented in this study have laid the groundwork for future research in the sports shooting domain.

Keywords: Machine learning, explainable AI, transformers, skeleton graphs, rifle shooting

## 1. Introduction

There are many factors affecting a shooter's performance in rifle shooting. Important factors are time spent in the aiming process, weapon movement before triggering, and body postural sway as a result of poor balance [1]. Because the eyes are focused on the target at the moment of shooting, they can not be used to control postural stability [2]. Several studies correlate poor posture control with poor shooting results for rifle shooting [3–6]. These studies mainly define posture control as body sway calculated from the force exerted by each foot on the ground measured with force plates. Because only balance in the legs is measured, these force plate sensors may not produce the full picture for the entire body's postural control. The force plates also restrict the subject to a stationary position. Therefore, it is interesting to examine other approaches that could be taken to gain a more comprehensive understanding of how postural stability affects shooting performance.

---

<sup>1</sup>This article received a correction notice (Erratum) with the reference: 10.3233/IDA-230950, available at [doi.org/10.3233/IDA-230950](https://doi.org/10.3233/IDA-230950).

\*Corresponding author: Jonatan Flyckt, Department of Computing, Jönköping University, Gjuterigatan 5 Jönköping, Sweden. Tel.: +46 73 687 53 26; E-mail: [jonatan.flyckt@gmail.com](mailto:jonatan.flyckt@gmail.com).

Sports shooting can be broadly divided into two domains: static shooting and dynamic shooting. This study focuses on dynamic shooting, which has a higher level of complexity with movements of both the shooter and targets. Both static and dynamic rifle shooting training is commonly conducted with one or multiple practitioners being instructed by a supervisor. For novices, training is mostly focused on striking static targets at fixed distances, where a supervisor focuses on factors such as stability, aiming, control, and movement [7]. Different shooting instructors can give contradicting feedback due to personal biases in the interpretation of data [8]. Consequently, there is a need for an objective and consistent data-driven feedback based on statistical analysis of data from real shooting scenarios.

Bio-mechanic pose estimation has been researched for almost half a century, often by using physical markers placed on human participants to build 3D body representations from visual sensors [9]. New techniques developed during the last decade allow for the tracking of joints by using image processing and machine learning. These technologies can build accurate 3D depth maps, producing so called *skeletons* of interconnected body joints. A relatively low-cost commercial sensor in this area is the *Kinect* body tracking sensor [10], which uses image and depth data with machine learning to build robust, view invariant skeleton models that measure 3D trajectories of human skeleton joints over time [10]. The latest version: *Azure Kinect*, is available since 2019.

As a result of these accurate low-cost technologies, the use of skeleton data together with machine learning is studied more actively, especially in the area of action recognition [11–15]. Skeleton data can also successfully measure motor functions [16], postural stability [17–19], as well as to some extent assess skill level in sports, e.g. handball [20]. There is a scientific consensus that postural balance is an important factor for rifle shooting performance [2–4,6,21]. Therefore, it is interesting to examine whether there are other postural factors besides postural sway that affect shooting performance, as such knowledge could be of assistance during the training of novice shooters.

### 1.1. Aim and scope

There is a research gap in machine learning approaches for skeleton data that explain their reasoning, as well as in using machine learning for decision support within the sports shooting domain. This paper presents a fully encompassing workflow ranging from data collection in live shooting tasks, to sensor merging, data processing, and the construction of explainable machine learning models. The *attention* mechanism in the deep learning *transformer* architecture is used to gain insights into what the model focuses on when it makes certain predictions [22]. The study will attempt to answer the research question:

*Can relevant factors for rifle shooting tasks be determined through posture and body movements?*

It is expected that, as a result of the proposed work, an increased understanding is gained of how to model experiments for shooting scenarios that use multiple body tracking sensors. The study will identify the possibilities and limitations of the gathered skeleton data, as well as give an initial insight into how skeleton data from live shooting scenarios can be modelled to generate relevant explainable factors.

## 2. Related work

### 2.1. Postural stability and shooting performance

Several studies have used force plate sensors to measure posture control in shooting tasks. However, these force plates are prohibitively expensive and can be impractical in situations outside controlled research studies and clinical environments [23]. The studies that have examined body tracking in shooting

tasks have had a delimited approach, mainly focusing on correlating shooting performance to mean sway velocity calculated from joint angles [5]. Skilled and experienced shooters have generally shown a better postural stability than novices [2–4]. Gun barrel stability has also been correlated to postural stability by several studies [3–5]. Although some studies have correlated posture to shooting performance between individuals [4], other studies have only shown intra-individual correlations [6]. These conflicting results, as well as a general lack of studies examining the effects of the posture and movements of the entire body on shooting performance, motivate further study.

## 2.2. Kinect and body tracking in scientific research

The Azure Kinect sensor [10] is a viable low-cost alternative to classical marker-based motion capture systems [19] as well as force plates [18] for measuring balance. Compared to motion capture suits, the Kinect has performed moderately to excellently depending on the tasks. The upside is that the participants are not restricted by wearing a motion capture suit. However, poorer tracking of the feet and ankles has been observed [16,19]. The Kinect sensor holds up well as a balance measurement when compared to force plate sensors [18,19], sometimes even outperforming them, with the upside of not limiting subjects to a stationary position. Downsides are limitations in accurately representing anterior-posterior swaying movements, overestimating large swaying movements, and having a framerate that is limiting for certain tasks [24].

Despite an exhaustive literature search, very few studies were found that use skeleton data to assess skill levels in humans. Most such studies have focused on isolated aspects of the body, such as angles between key joints in handball throws [20], or the take-off velocity of jumping motions [24].

Because of occluded joints forcing the Kinect sensor to estimate positions, Gao et al. [25] found that using two Kinect sensors placed in front of the target at an angle led to higher body tracking accuracy. Núñez et al. [26] proposed data augmentations on the skeleton data by shifting the graph slightly, because limited amounts of data would cause generalisability issues for machine learning models. Vemulapalli et al. [27] suggested that model performance could be improved by making the skeletons invariant to the absolute locations of subjects. They did this by transforming the coordinates to a body-centric view with the hip center as the origin, and the x-axis running parallel to the hip.

## 2.3. Action classification

Action classification tasks use time-series skeleton data, often with deep learning techniques, to classify human actions such as jumping, running, throwing etc. This section provides a brief chronological survey of the area.

Du et al. [11] used a hierarchically bi-directional RNN approach, and found that it was beneficial to divide the skeleton into five body parts (arms, legs, torso) before fusing them iteratively into one feature representation. Liu et al. [28] achieved good results with Long Short Term Memory (LSTM) and hierarchical tree traversal techniques, in large part due to the LSTM's inherent strengths in discarding irrelevant information such as weak sensor data. Adding the attention mechanism to this approach improved the LSTM's global context memory and increased predictive performance [29]. Attention has also been able to locate the most discriminative time frames across longer sequences of skeletons, e.g. being able to identify the point at which the arm begins to stretch in a punching motion [30].

Ke et al. [12] used convolutional neural networks (CNNs) to classify actions with skeleton data, arguing that they could remember long time sequences better than LSTMs. They restructured skeleton sequences into three cylindrical coordinate channels, each composed of four frames. Convolutions on graphs have

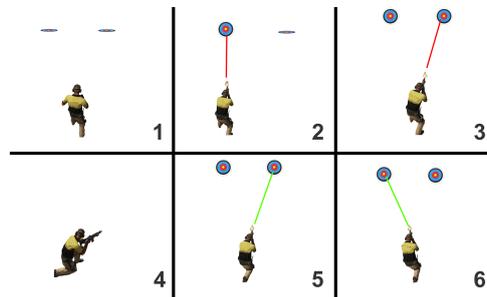


Fig. 1. Scenario illustration. Sequential illustration of the shooting scenario for the data collection. Source: [33].

also been used to automatically extract feature maps from joints connected spatially between each other, as well as temporally through time [15]. Plizzari et al. [31] combined a high level local feature extracting Graph Convolutional Network (GCN) with transformers and self-attention, allowing for explanations of what features were mutually important for predictions, as well as alleviating the long temporal relation issues of LSTMs [22]. Similarly to Song et al. [30], they separated the temporal and spatial skeleton graphs into separate streams, which allowed the model to attend both to the important time frames and the joints with the most discriminative power. Transformers are composed solely of attention layers, and have an added benefit compared to other models of being able to explain their reasoning using the attention matrices produced from each attention layer [22].

### 3. Method

#### 3.1. Experiment

The research question will be answered with the help of an experiment which will rely on data gathered from several human participants. The participants will perform a rifle shooting task, as described in the data collection section. The experiment is a quasi-experiment, meaning that there will be no random assignment, and that the potentially discovered relations in the data may not be true cause and effect relations [32]. The experiment will be controlled in the sense that there will be a full control over the manipulation of the data, and as many iterations as needed can be performed. The dependent variables will be the various shooting poses of participants as well as which orientation (strong or weak hand) the participants use. The independent variables will be the sequential body tracking data produced by the Azure Kinect sensors.

#### 3.2. Data collection

The data collection for this study relied on one well-defined scenario, with a focus on identifying the postural effects on shooting performance with a semi-automatic rifle. The data collection was performed in an indoor shooting range to limit the effects of weather and wind on the participants and sensors. The scenario was designed to measure shooting performance when switching between different shooting targets and body postures (Fig. 1).

##### Scenario description:

1. The participant is equipped with a semi-automatic rifle (HK416 with a 5.56 mm × 45 calibre) (Fig. 2).



Fig. 2. HK416. The HK416 semi-automatic rifle (5.56 mm × 45 calibre) used in the data collection scenario.

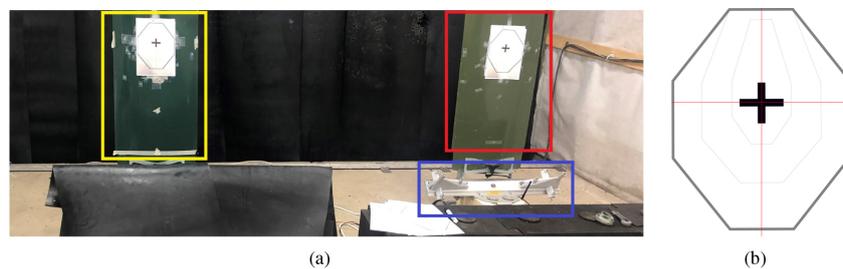


Fig. 3. Targets and shot detection system. (a): The yellow and red squares show the left and right pop-up targets used in the data collection. The blue square shows the LOMAH system [34], responsible for detecting shot positions. (b): The International Practical Shooting Confederation (IPSC) target plate with scoring zones that was used in the data collection, the size of an A4 paper sheet.

2. Two pop-up targets are placed 20 metres away from the shooter, with roughly 1.5 metres between each other (Fig. 3).
3. The participant starts in a standing, shooting ready position.
4. The left target pops up, the participant aims at the target and fires a set of three shots.
5. The participant switches back to a shooting ready position.
6. 15 seconds after the left target popped up, the right target pops up. The participant switches targets and fires a set of three shots at the right target.
7. The participant switches to a kneeling position.
8. The participant fires another set of three shots at the right target.
9. The participant switches targets and fires a final set of three shots at the left target.
10. The participant secures and unloads the weapon.

Figure 3b shows the shooting target that was used in the data collection scenario. The participants were instructed to hit as close to the centre as possible while still maintaining a high speed of execution.

In order to build a robust and diverse dataset it would have been ideal to have shooting participants of different experience levels. However, due to safety constraints only experienced shooters with official weapons training could participate. 13 shooters were used in the data collection, 12 of whom had a military background, three of whom had a sports shooting background, and three of whom had a background in hunting. All participants were male, and ranged from the age of 31 to 62, with an average age of 48. In an attempt to simulate poor posture, the participants performed half of the scenarios with their weak hand. Each participant performed the scenario six times, resulting in a total number of 78 recorded shooting scenarios. Figure 4a,b shows a participant performing the scenario.

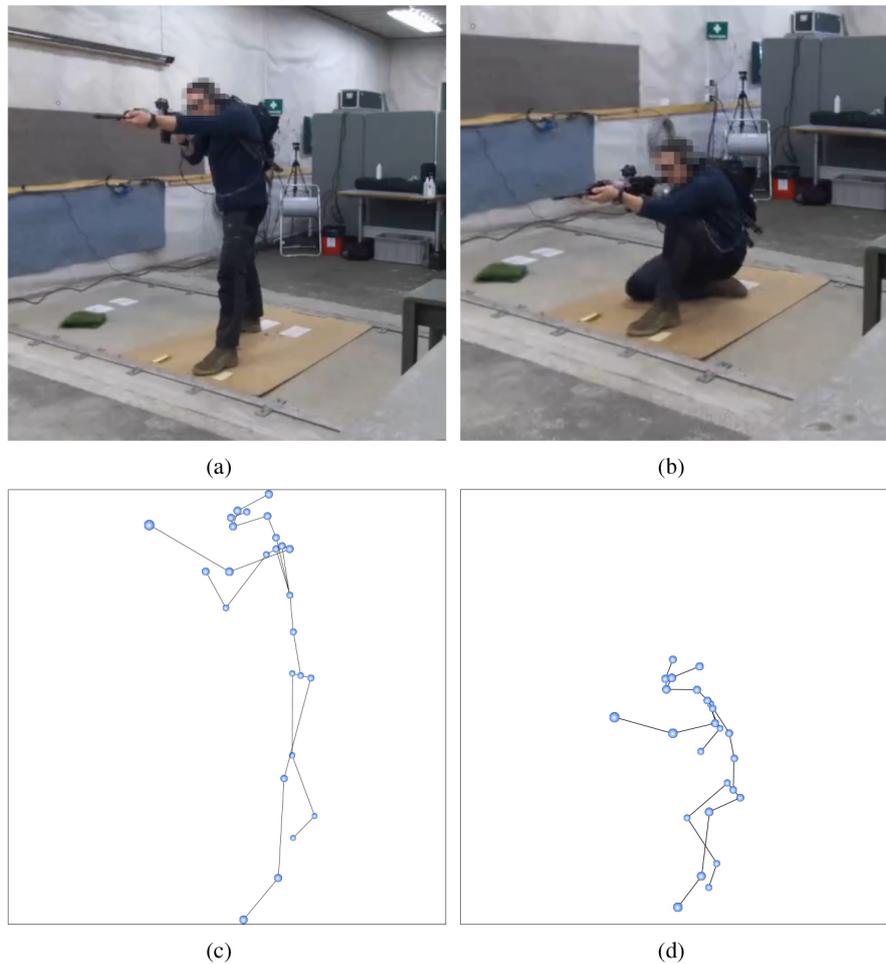


Fig. 4. Data collection participant and resulting skeletons. (a,b): A participant performing the data collection scenario, first standing (a) and then kneeling (b). (c,d): The resulting skeleton representations.

The Azure Kinect body tracking sensor [10] was used to capture the body movements of the participants during the scenario. The accuracy of the Kinect sensor can be affected by several factors such as occlusion of body parts, e.g. a hand behind the body from the view of the camera. Other factors include poor lighting conditions, or disruption of the sensors' infrared signals from e.g. sunlight or infrared heaters. To mitigate the effects of poor sensor readings, three Kinect sensors were used: one behind the participant, and two to the front on either side of the participant at roughly a 45 degree angle, as suggested by Gao et al. [25]. The sensor behind the participant was placed slightly to the right, because when placed straight behind the participant the sensor would have difficulties with determining in which direction the participant was facing. To ensure a consistent and coordinated trigger timing of 30 frames per second on all three devices, the Kinect sensors were connected together with a synchronisation wire signalling when to capture new frames.

The Location of Miss and Hit (LOMAH) system (Fig. 3a) was used to record the time and position of shots by detecting the sound waves produced by bullets [34]. These shot positions were recorded in an external system as horizontal and vertical distance in millimetres from a defined centre of target position.

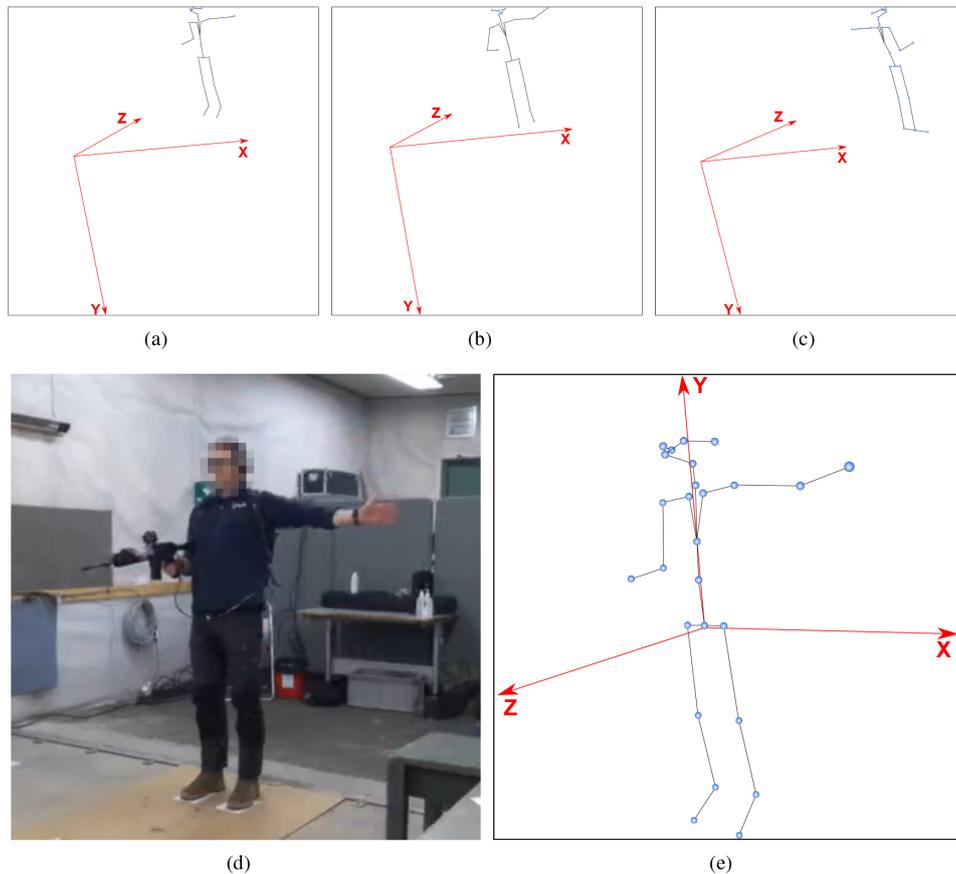


Fig. 5. Skeleton synchronisation and merging. (a–c): A participant performing the synchronisation movement from the views of the front left (a), front right (b), and back (c) sensors, with the origins of their coordinate systems at the sensor positions. (d): A captured frame of the participant from the view of the front left sensor. (e): The merged version of the sensors, with the coordinate system aligned with the body positions.

A microphone worn on a backpack equipped on the participant detected shots and ensured that the Kinect data could be matched to the shots detected by the LOMAH system.

The large computational resource requirements of the Kinect sensors necessitated the use of two different machines to process the data, in order to retain the highest possible frame rate (30 hertz). This meant that the Kinect sensors were unaware of each other's time systems, making it difficult to match which frames represented the same movements. Consequently, each participant was instructed to perform a synchronisation movement before each scenario by extending their non-shooting arm upwards and moving it slowly outwards in an arc away from the body and down to the side (Fig. 5).

### 3.3. Data pre-processing

The Kinect produced 32 skeleton joints for each person present in each frame (roughly 30 frames per second). The joints broadly represented physical joints in a human body, e.g. hips, shoulders, knees etc. Because hand joints received poor estimations when further than 1.5 metres away from the sensor, they were excluded from the data. For each joint the following variables were collected:

- X position in millimetres from the sensor, horizontally from the view of the sensor.
- Y position in millimetres from the sensor, vertically from the view of the sensor.
- Z positions in millimetres from the sensor, extending straight out from the sensor.
- Confidence level for the joint (specifying whether the joint was in view of, occluded from, or too distant from the sensor).

Because the raw data from the three Kinect sensors were expressed in their own absolute cartesian coordinate system with the origin located at each sensor's position (Fig. 5a-c), some pre-processing was required. The first steps involved matching the raw data from the three devices both temporally and spatially, and to isolate the body of interest, i.e., the shooter. A body-centric coordinate system was used, as suggested by Vemulapalli et al. [27]. The synchronisation movement performed by the participant was used to synchronise sensor data temporally across Kinect devices, and as a frame of reference for calculating new unit vectors for each device. The synchronisation skeleton frame index ( $S$ ) and the following ten time frames (roughly 0.33 seconds) were used to calculate the new unit vectors, from which transformation matrices were constructed for each sensor device's skeleton data. The normalised unit vector  $\hat{u}$  (representing the new X axis) was defined as the average vector position between the left hip ( $\vec{h}_l$ ) and right hip ( $\vec{h}_r$ ) joints during the synchronisation sequence Eqs (1) and (2).

$$\vec{u} = \left( \sum_{i=S}^{S+10} \vec{h}_{li} - \vec{h}_{ri} \right) \quad (1)$$

$$\hat{u} = \frac{\vec{u}}{|\vec{u}|} \quad (2)$$

The normalised unit vector  $\hat{v}$  (representing the new Y axis) was defined as the average position of the vector formed from the spine-chest ( $\vec{g}_c$ ) joint and the pelvis ( $\vec{g}_p$ ) joint during the synchronisation sequence Eqs (3) and (4).

$$\vec{v} = \left( \sum_{i=S}^{S+10} \vec{g}_{ci} - \vec{g}_{pi} \right) \quad (3)$$

$$\hat{v} = \frac{\vec{v}}{|\vec{v}|} \quad (4)$$

The unit vector  $\hat{w}$  (representing the new Z axis) was defined as the normalised cross product of the new X and Y axes Eqs (5) and (6).

$$\vec{w} = \hat{u} \times \hat{v} \quad (5)$$

$$\hat{w} = \frac{\vec{w}}{|\vec{w}|} \quad (6)$$

The origin  $\vec{O}$  used to translate coordinates to the new system was defined as the average pelvis ( $\vec{g}_p$ ) joint position during the synchronisation sequence (Eq. (7)).

$$\vec{O} = \frac{1}{N} \sum_{i=S}^{S+10} \vec{g}_{pi} \quad (7)$$

Each joint vector  $\vec{a}_i$  from the entire skeleton sequence was then transformed into  $\vec{b}_i$  by computing the dot product of the translated joint position and the change-of-basis matrix. The change-of-basis matrix

was constructed by appending the columns of the three unit vectors and taking the inverse of the resulting matrix Eq. (8).

$$\vec{b}_i = \begin{bmatrix} \hat{u}_x & \hat{v}_x & \hat{w}_x \\ \hat{u}_y & \hat{v}_y & \hat{w}_y \\ \hat{u}_z & \hat{v}_z & \hat{w}_z \end{bmatrix}^{-1} \cdot \begin{bmatrix} a_{ix} - O_x \\ a_{iy} - O_y \\ a_{iz} - O_z \end{bmatrix} \quad (8)$$

Despite the sensor data being transformed into the same coordinate system, the joint estimations and body rotations from the three sensors differed slightly, which made skeleton merging difficult. To overcome this, the sensor on the side of the weapon-holding hand was used as a reference to attach body-centric joint positions from the other sensors to this sensor's skeleton. These joint positions were calculated by performing new body-centric transformations on each skeleton frame for each sensor, ensuring that the body orientation was not affected by the differing sensor estimations. All joints with a high confidence value (i.e. not occluded by other body parts from the view of a sensor) were used to compute an average joint position  $\vec{c}_i$ . These average positions were then transformed and translated into  $\vec{d}_i$ , to be represented in the coordinate system of the sensor on the side of the weapon-holding hand, using this sensor's origin  $\vec{O}$  and change-of-basis matrix  $[\hat{u}, \hat{v}, \hat{w}]$  Eq. (9).

$$\vec{d}_i = \left( \begin{bmatrix} \hat{u}_x & \hat{v}_x & \hat{w}_x \\ \hat{u}_y & \hat{v}_y & \hat{w}_y \\ \hat{u}_z & \hat{v}_z & \hat{w}_z \end{bmatrix} \cdot \begin{bmatrix} c_{ix} \\ c_{iy} \\ c_{iz} \end{bmatrix} \right) + \begin{bmatrix} O_x \\ O_y \\ O_z \end{bmatrix} \quad (9)$$

Although three sensors were used in order to prevent the occlusion of body parts, joints were sometimes out of view from any of the sensors, resulting in outlier positions. To simulate the actual trajectory of these joints, their positions were estimated with a loosely fit 4th degree polynomial regression from the surrounding high confidence frames. To remove twitching joint movements, a median smoothing was performed on each frame from the surrounding five frames, followed by a mean smoothing from the surrounding three frames, resulting in a smooth merged skeleton representation (Fig. 5e). The shots detected by the LOMAH system were matched temporally to the skeleton data by detecting the shots with a microphone. Any samples containing shots that the microphone sensor failed to detect were discarded.

Features were extracted from the merged skeletons based on the angles of *bones* formed by the vectors between connected joints (e.g. the right femur bone was defined as the vector between the right hip and knee). This representation retained most of the information in the data, but made it invariant to the absolute position of the body in a coordinate system, allowing the model training on the data to focus on finding relevant patterns. The angle values for a bone were represented in three radian values defined by the angle value of the bone against the X, Y, and Z basis vector respectively (Fig. 6).

To make the models more robust, a set of data augmentations were used: For each skeleton sequence, six new orientations were produced from rotations around the Y axis with 10, 20, and 40 degrees of rotation to both the left and the right, and used to train the model in addition to the original orientation, as suggested by Núñez et al. [26]. Gaussian noise was also added randomly during training to some of the bone features in an attempt to make the models more robust to imperfect data.

### 3.4. Feature embedding and ViT adaptation

Because of the strength of Transformers on sequential data tasks, as well as their inherent explainability capabilities [22], a Vision Transformer (ViT) was chosen for the learning tasks. ViTs divide images into equally sized rectangular patches, and embed them into projected one-dimensional vector representations by feeding them through trainable embedding layers. These embeddings are then treated in the same

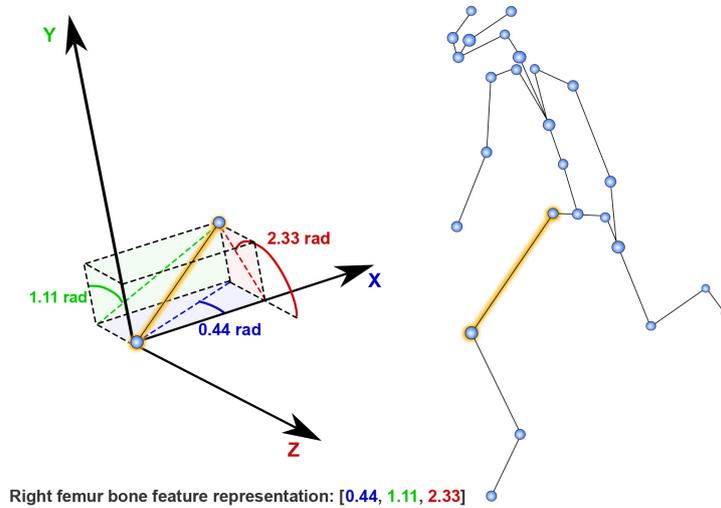


Fig. 6. Skeleton bone feature representation. Each bone was defined by three features per frame (X, Y, and Z angles), here illustrated for the right femur bone from a frame in a skeleton sequence.

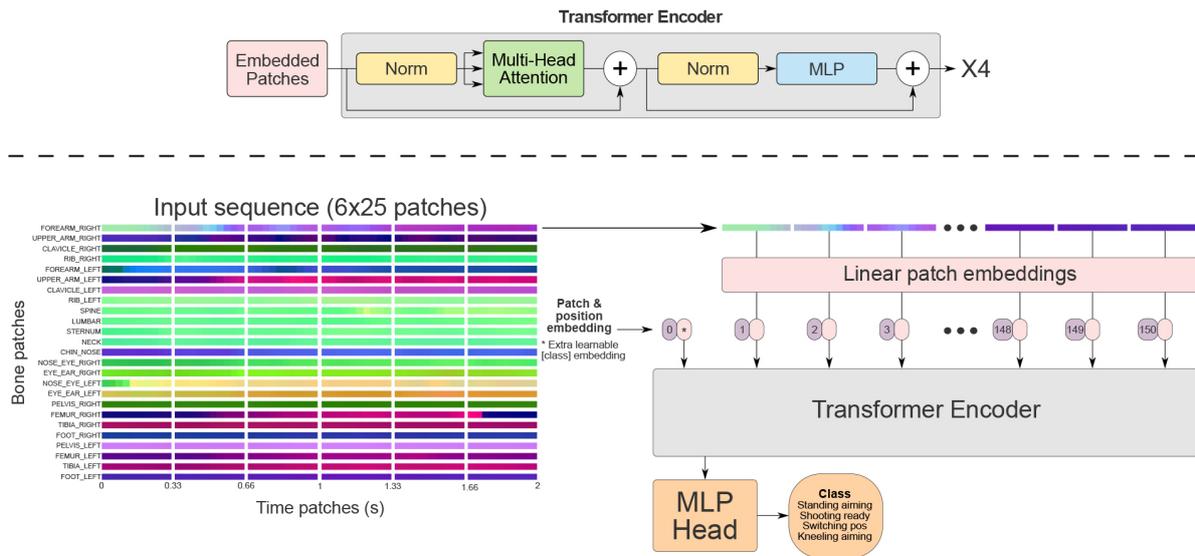


Fig. 7. Model architecture. The input patches produced from a skeleton sequence, and how they were processed by the ViT model. Adapted from Dosovitskiy et al. [35].

way as positional tokens in standard transformers. This means that ViTs are not limited solely to use on images, but can be used on any data that can be represented with patch embeddings. ViTs can explain their predictions by using the attention produced from the patch embeddings [35].

Each skeleton sequence was represented with time frames as columns, and the 25 different bones as rows with the X, Y, and Z radian values stacked on each other in a fixed feature representation of  $3 \times 25 \times 60$  scalar values, i.e. a skeleton sequence of two seconds with 30 frames per second, and embedded this into patches of size  $3 \times 1 \times 10$ . This meant that each patch represented one bone and its radian X, Y, and Z components over a time period of roughly one third of a second (Fig. 7). The feature representation was

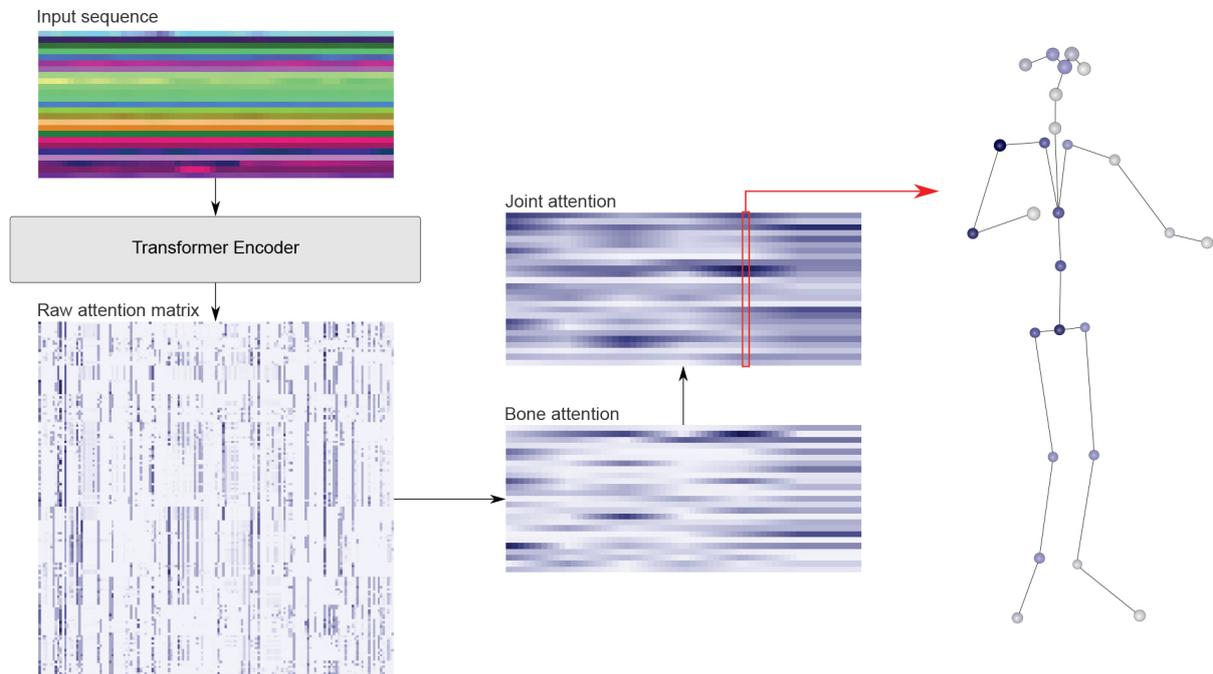


Fig. 8. Attention extraction. Pipeline describing the process from an input skeleton sequence to an explainable attention visualisation for a model prediction.

partly inspired by Ke et al. [12], who divided their data into time clips consisting of separated cylindrical coordinates. Because the patches were one row in height, representing only one bone, each bone for a moment of time could attend to any bone and time (including itself) across the patch embeddings. This allowed for the extraction of the attention for each bone and sub-sequence for a prediction, explaining some of the model's reasoning. The ViT model used four multi-head attention layers composed of eight heads for each layer. A multilayer perceptron (MLP) head attached at the end of the model produced the output(s) for the specific task (Fig. 7). Each attention layer used a Gaussian Error Linear Unit (GELU) activation function. For both learning tasks, a dropout of 0.3 was used, as well as a reduce-on-plateau learning rate starting on  $1 \times 10^{-3}$ , and reducing by a factor of 0.3 after plateauing for three epochs.

For each prediction, the model produced 32 attention matrices from the eight heads of the four attention layers, which were summed up into one representation of the total attention. Figure 8 shows how the attention map was processed from the raw attention map produced by the model into a format that was deemed easier for humans to interpret, explaining which joints affected the prediction for each time frame.

### 3.5. Shooting pose estimation

A pose classification task was used for demonstrating the capabilities of the workflow presented in this study. Four pose classes were identified and labelled from the data collection scenario: *standing shooting ready*, *standing aiming*, *switching position*, and *kneeling shooting*. Estimating the pose of a shooter in a shooting scenario could be useful in a live prediction situation in a larger system. Such a system could potentially be made up of many different models trained for judging different shooting poses, and would therefore need some way of determining the current shooting pose. The labels were

Table 1  
Class distribution for the pose estimation task

Pose class	Samples	% of distribution
Standing shooting ready	6,976	43.39
Standing aiming	2,329	14.48
Switching position	1,864	11.59
Kneeling aiming	4,910	30.54

Table 2  
Class distribution for the shooting hand estimation task

Class	Samples	% of distribution
Strong hand	3,700	49.47
Weak hand	3,779	50.53

automatically assigned based on the timing of shots. The frames from 0.5 seconds before the first shot of a 3-shot sequence to the last shot of that sequence were labelled as *standing aiming* or *kneeling aiming*. Each 60 frame (two second) long skeleton sequence belonging entirely to one class label was extracted from the scenario. New sequences were extracted starting at every third frame. For this learning task, a cross entropy loss over four output nodes for the four classes was used, as well as a batch size of 256 samples. Because the classes were not equally distributed (Table 1), an imbalanced class data sampler was used to semi-randomly upsample the classes with few samples, and downsample the classes with many samples during training. The 13 participants were separated into 13 folds and used in a non-random repeated cross-validation. To obtain robust result, we trained 12 different models for each participant, where the samples from one other participant acted as the validation set one time each, and the remaining 11 participants acted as the training set. This resulted in a total of 156 models being trained for the experiment. We kept participants entirely in either test, validation or training to ensure that the model did not overfit to one specific participant during training. The training participant samples were fully augmented as described in the Section 3.3.

### 3.6. Shooting hand and shooting performance estimation

Due to the relatively low amount of participants and data samples, we could not predict skill through continuously increasing skill levels in a regression task with the shooting score as the target variable. Instead, a binary classification task was constructed to test whether our approach could separate skilled and unskilled shooters from each other. The samples produced from the scenarios where the participants used their strong shooting hand were used to denote good posture, and the samples that used the weak hand were used to denote poor posture. The reasoning was that each shooter was much more accustomed to shooting with their strong hand, and would therefore make mistakes when shooting with their weak hand. The shooting accuracy and speed differences between strong and weak hand shooting showed that this was the case for most of our participants; with the exception of one participant the performance became worse when using the weak hand. The average time taken for a three shot series with the strong hand was 2.6 seconds with an average shot score of 2.19, whereas it took 3.4 seconds with an average shot score of 1.89 with the weak hand. Those who had a more similar performance between their strong and weak hand were also observed to have a more similar posture than those who had a large performance difference when using different shooting hands.

The same feature representation, model architecture, and non-random repeated cross-validation were used as in the pose estimation task (13 folds, 156 models). To limit the scope of this task, only standing

Table 3  
Pose estimation metrics

Metric	Total <sup>a</sup>	CI 95% <sup>b</sup>
Accuracy %	80.00	[78.55 , 81.73]
$\kappa$ rating	0.713	[0.688 , 0.729]

<sup>a</sup>The results of the predictions from all participants' samples. <sup>b</sup>Confidence intervals at 95% confidence level.

Table 4  
Pose estimation attention statistics

Bone	Average attention ranking <sup>a</sup>
FOREARM_RIGHT	1.15
FOREARM_LEFT	2.15
UPPER_ARM_RIGHT	3.62
FEMUR_LEFT	4.15
FEMUR_RIGHT	6.92
PELVIS_LEFT	7.62
UPPER_ARM_LEFT	8.15
EYE_EAR_LEFT	8.69
NOSE_EYE_LEFT	9.38
PELVIS_RIGHT	9.69

Shows the top 10 (out of 25) bones for predicting poses. Note that left-handed samples were mirrored, and thus FOREARM\_RIGHT denotes the forearm of the arm that pulls the trigger for both left- and right-handed samples. <sup>a</sup>The average attention ranking calculated from the 13 individual attention rankings produced from each participant.

shots were investigated. We extracted two-second samples by looking at the frames from 1.5 seconds before the first shot of a standing three shot series to 1.5 seconds after the last shot in the same series, extracting new sequences starting at every third frame. The class distribution was very close to equal between the two classes (Table 2). All skeletons were adjusted to a right-handed orientation by mirroring the left-handed shooting scenarios against the plane formed by the Y and Z axes, and switching the positions of the joints on the left and right side of the body. This ensured that the model could not learn to simply classify left- and right-handed scenarios, but that it would have to learn relevant factors of the actual poor posture that was produced from the shooters using their weak hand.

## 4. Results and analysis

### 4.1. Pose estimation

Table 3 shows the total results for the pose estimation task in the form of total results and the 95% confidence intervals based on the results from each fold (participant) in the experiment. Because the class distribution was imbalanced, both Accuracy and Cohen's  $\kappa$  were used as evaluation metrics, as Accuracy would be biased towards majority classes, whereas Cohen's  $\kappa$  measures how much better the prediction was than a random prediction given the class distributions [36]. It is worth noting that the confidence intervals were computed with unequal fold sizes; the participant with the most samples had 1792 samples, whereas the one with the fewest samples had only 815 samples. The model performed best on *standing shooting ready* and *kneeling aiming*, but often confused *switching position* with *kneeling aiming*, as well

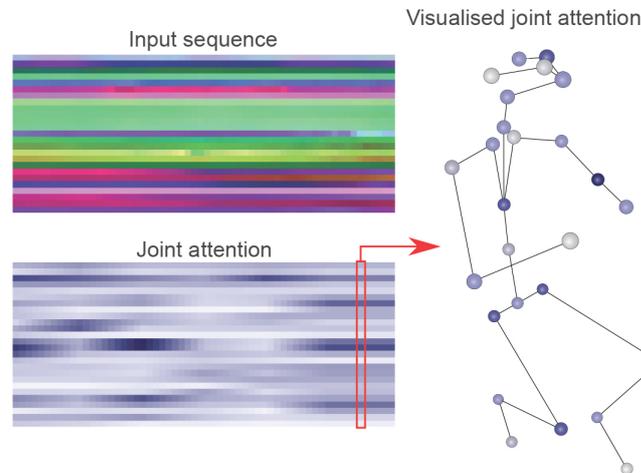


Fig. 9. Attention visualisation: Switching position. The input skeleton sequence of one of the participant switching from a standing to a kneeling position, the joint attention produced from the attention matrix from the pose estimation model, and a skeleton visualisation of the joint attentions for one time frame. Darker colour indicates stronger attention.

as *standing aiming* with *standing shooting ready* to some extent. These were reasonable errors, as these were the classes most similar to each other. In addition to this, the automatic labelling was not always 100% correct.

Attention statistics were calculated for the model by ranking the importance of each bone feature for each fold (participant), and then computing the average of these rankings across all participants. The attention often focused on the position of the arms or the femur bones (Table 4), which was reasonable, as the angle of the arms could help determine whether a participant was aiming or not, and the angles of the femur bones could help the model determine whether a participant was standing or kneeling. Figure 9 (*switching position*) shows how the processed attention produced from the model was visualised on a 3D skeleton model. This figure highlights how the model attended to the left femur bone in the visualised time frame, but also earlier in the sequence when the participant was in a standing position. This indicates that the model has learned to attend to the changing angle of the femur bones to determine whether the sequence involved switching from a standing to a kneeling position or not.

#### 4.2. Shooting hand and shooting performance estimation

Table 5 shows the results from the shooting hand estimation experiment. Overall, the results indicate that it is possible to some extent to separate participants using their strong hand from participants using their weak hand. However, a large difference was observed in prediction performance between the participants, as well as for each individual shooter depending on which other participant acted as validation data and thus stopping criterion for the training. It is not entirely clear why there was such a big difference in model performance between participants, but most likely more training data is needed. Another reason for the difference in performance could be the differing shooting styles, and too many pose faults in relation to the number of collected samples. The model may have learned simple indicators of good versus bad posture, but these could be difficult to generalise between shooters with e.g. aggressive or more relaxed poses.

Table 6 shows the computed attention statistics from all trained models for the shooting hand estimation task. These statistics indicate that the model has learned that the main difference in posture for strong

Table 5  
Shooting hand estimation metrics

Metric	Total <sup>a</sup>	CI 95% <sup>b</sup>
Accuracy %	71.56	[66.88 , 73.52]
$\kappa$ rating	0.431	[0.332 , 0.462]

<sup>a</sup>The results of the predictions from all participants' samples. <sup>b</sup>Confidence intervals at 95% confidence level.

Table 6  
Shooting hand estimation attention statistics

Bone	Average attention ranking <sup>a</sup>
FOREARM_RIGHT	1.00
NOSE_EYE_LEFT	3.92
LUMBAR	5.15
EYE_EAR_RIGHT	5.38
STERNUM	6.62
SPINE	8.54
CLAVICLE_RIGHT	8.92
CLAVICLE_LEFT	9.15
RIB_LEFT	9.31
EYE_EAR_LEFT	9.31

Shows the top 10 (out of 25) bones for predicting whether shooting with the strong or weak hand. Note that left-handed samples were mirrored, and thus FOREARM\_RIGHT denotes the forearm of the arm that pulls the trigger for both left- and right-handed samples. <sup>a</sup>The average attention ranking calculated from the 13 individual attention rankings produced from each participant.

versus weak hand shooting lies in the position of the forearm of the trigger arm. Other important features came from the head and centre torso body parts. This is in line with observations that were made from studying skeleton samples, where the shooting samples using the weak hand often had a more unnatural pose where the participants leaned into the weapon differently. Figure 10 highlights the difference in pose for one participant while actively firing the weapon during the scenario. This figure shows that the elbow of the trigger arm is raised when using the strong hand, whereas the participant is more huddled together and leaning unnaturally into the weapon with the weak hand. This is in line with many of the other top attended features from Table 6. Overall, domain experts noted an unrelaxed shooting position of the head and shoulders when studying individual samples using the weak shooting hand.

An attempt was also made to predict skill by modelling a regression task where the target variable was the shooting score achieved by each participant, but this yielded close to random predictions. Most likely a lot more data is needed for such a task.

## 5. Discussion

This study could correlate shooting performance to posture to an extent by using strong and weak hand scenarios to denote good versus bad posture (Table 5). However, the results varied widely both between different participants, as well as between which validation fold was used as stopping criterion for the training. It is unclear why such a difference in performance between participants was produced; more

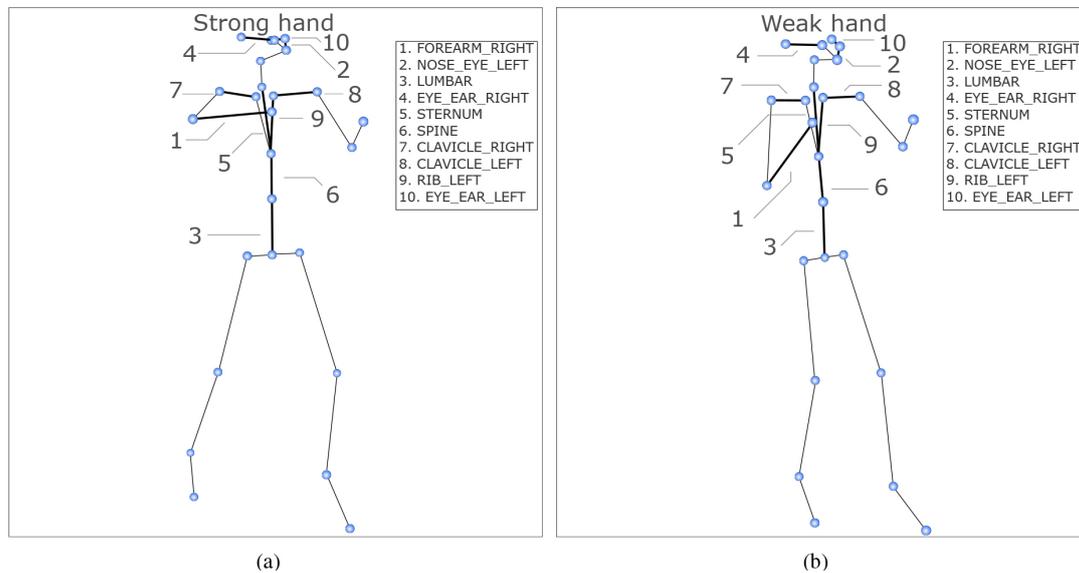


Fig. 10. Strong and weak hand comparison. Shows a comparison of one of the participants in the act of firing their weapon with their (a) strong hand and (b) weak hand. Highlighted are the top 10 bones from the overall attention statistics for the shooting hand task (Table 6). Note that all left-handed samples were swapped to a right-handed orientation.

experiments are needed. Because the experiment shows such volatility, it is likely that more training data is needed for the models to be more stable and statistically significant. There are also other factors that could affect the performance, such as poor sensor readings. Additionally, body postures in general may not have a large enough effect on shooting performance to be used as a sole factor; many other factors such as trigger jerking, small weapon movements, and eye movements could potentially have a larger impact on performance [6]. There have also been conflicting reports on whether postural stability correlates to shooting performance across multiple individuals [4,6]. Various shooting styles were observed from the participants; some had a very forward leaning, aggressive shooting pose, whereas others had a more upright pose while still achieving high scores. With only 13 participants, these differences in body postures may have been detrimental to the discriminative powers of the model. The results that were attained from estimating shooting performance through shooting score as a regression target were essentially random. For the regression approach to work it is likely that much more training data is needed. It is also possible that using shooting score as a regression target is simply not a good approach to predict skill. A potential future approach could be to use a combination of simulated poor posture from using the weak hand, expert judgement of shooter posture, and a wide variety of skill level between shooters to bin participants into different levels of posture quality. A  $\kappa$  rating of 0.431 (Table 5) for the shooting hand task indicates that the area deserves further examination, and that it has promise for future studies.

The pose estimation task saw stronger performance (Table 3), likely due to the more apparent differences in poses in different phases of the shooting scenario compared to the differences between good and poor posture. The model learned a reasonable representation, as the poses that were classified incorrectly were usually the most similar to each other. Because the pose labels were produced automatically by using shot moments, the labels were not always accurate. All frames up until 0.5 seconds before the first kneeling shot were labelled as belonging to the *switching position* class. However, many participants spent a significantly longer time after switching position before firing their first shot, meaning that more frames should have been labelled as *kneeling aiming*. Many of the standing shots were also incorrectly

labelled, as many participants continued to aim for some time after firing their first standing shots instead of moving to a shooting ready position with the weapon aimed down. Some participants also aimed for more than 0.5 seconds when standing up before firing their first shot. In fact, when studying individual predictions next to the samples, it was observed that the model's predictions were often better than the labels, which indicates that better results could probably be attained with better labelling. Using a combination of up- and downsampling seems to have removed any bias towards majority classes, despite the largest class having roughly four times more samples than the smallest class (Table 1). Taking into account the incorrect labels and the class imbalance, the model can be judged to have performed quite well. Manual study of the samples and attentions as well as the attention statistics in Table 4 indicate that the model has learned to look mostly at correct factors for discriminating between the classes.

The attention mechanism produced explainable predictions from the models, often being able to attend to the relevant frames and bones, as was also observed by Song et al. [30] and Plizzari et al. [31]. Because neither task saw excellent results, one also has to consider that a non-negligible part of the attention statistics were computed from incorrect predictions, and should therefore be seen as more of an indication of bone importance, than as an absolute truth. Figure 9 shows how a frame from a prediction could be explained through visualisation by colouring the joints of a skeleton frame with the values from a processed attention matrix. Longer sequences were visualised using 3D modelling or videos of skeletons with continuously shifting levels of attention on joints. It is the opinion of the authors that the visualisation provides an intuitive insight into the model's reasoning, which can aid both in better model development, and as a basis for decision making during rifle shooting training.

A limiting factor to the proposed transformer model was the fixed input size, i.e. all bones and their radian X, Y, and Z values over a sequence of 60 time frames (two seconds). Using longer sequences would require aggregating the data along the time dimension, which would cause the model to perceive subjects to be moving much faster than they were in reality. A potential workaround to this would be to include a time token in the input sequence, indicating to the model that it should judge samples differently based on time taken. Although standard transformer models could use different input sizes, longer sequences could possibly be more difficult to train on, and thus require more data. Other sequential models such as LSTMs, or GCNs using graphs across time could work well for the learning tasks, but would not produce the same level of explainability.

Some compromises were made in the selection of participants due to safety constraints, both in the number of participants and the amount of scenarios they could perform. Because of this, the data may be skewed to one type of behaviour. Ideally there would have been a wide variety of skill levels represented among the participants. Having half the scenarios be performed with the weak shooting hand to simulate different skill levels worked quite well to increase the diversity in the dataset; shooters shot better and faster with their strong hand, and their body postures and movements were generally more in line with shooting doctrine according to domain experts.

The use of three body tracking sensors to limit the poor pose estimations caused by occluded body parts worked to an extent. However, joints were still occasionally occluded, and would stutter substantially between frames, which may have limited the possibilities for the machine learning models to focus on small-scale details in shooting posture. Although standing poses were represented reasonably well, kneeling or sitting postures often had very poor joint estimations in the legs. Using polynomial regression to interpolate low confidence joints by looking at surrounding high-confidence frames of the same joints served as a good heuristic for actual joint positions, making the skeletons represent reality better. The synchronisation movement helped to match the three sensors' skeleton data temporally and spatially, although the poor sensor readings made it difficult to perform a perfect matching. Despite transforming

the different sensor data to the same coordinate system, their differences in intra-skeleton joint estimations was an additional issue for skeleton merging; one sensor would estimate a joint to be at a slightly different distance and angle from the other joints of the body than another sensor. Performing a transformation on each single frame by using one sensor's skeleton as the main skeleton helped to circumvent some of these issues.

### *5.1. Future work*

There are many possibilities for future studies to expand on the work in this study. For better shooting performance estimation, there may be a need for more manual features based on expert knowledge of the shooting domain, such as mean sway velocity and other established factors [2–4]. Although deep learning models can generally find important features from raw data, the relatively small dataset could benefit from handcrafted features with more discriminative power. Such features could be used both in classical machine learning models or as additional dimensions in the architecture proposed in this study. Another future step could be to examine different ways of modelling shooting skill through classification or regression tasks. Different input sequence lengths would allow for studying both long- and short-term movement patterns. This could be done by warping inputs of different lengths to one size and including some time indication token, or through using multiple models with different input sizes, or models capable of dealing with varying input sizes. It could also be interesting to examine whether a different coordinate representation (e.g. cylindrical [12] or spherical) could have an effect on model performance.

If shooting performance estimation was more successful in the future, the attention explanations could be used as a direct visual feedback to a novice shooter in a product used in live training scenarios by highlighting important aspects of their posture control. This study has examined and demonstrated one of the ways in which attention maps can be visualised through simple sums of the attention placed on a bone in time, but there are many additional possibilities for explaining predictions with visualisations. Because it is known which pair of time/bone patch attend to each other (or themselves), visualisations of the attention between pairs of bones/joints could extend the explanations further. Attention could also help during model development as a tool to identify incorrectly learned patterns, or where the features were lacking. It would also be interesting to perform user studies with shooting instructors to determine whether the attention visualisations were helpful and intuitive, and use the expert assessment to improve explanations further.

Additional data quality improvement techniques such as smoothing could be examined further, as the poor joint estimations produced by the sensors were potentially detrimental to model performance. Other body tracking approaches such as body tracking suits could potentially be an alternative in order to produce more accurate skeleton representations [16,19]. More data samples overall are probably needed, with increased diversity among shooter styles and skill levels. Different shooting scenarios with a bigger range of movements could also be of interest for future data collections.

## **6. Conclusions**

This study has examined postural factors in rifle shooting scenarios through the use of multi-sensor body tracking. The study has described the difficulties of large-scale data collection involving human participants and body tracking sensors, and the mitigating measures taken to produce high quality data. A system has been developed for merging data from multiple body tracking sensors with differing time and coordinate systems, and pre-processing steps to smooth out the data and represent it as input features

into machine learning algorithms. The approach to generate explainable predictions from multi-sensor body tracking is general, and can be adapted to domains other than rifle shooting. It can be argued that relevant factors for rifle shooting tasks have been extracted from postures and body movements to an extent; the shooting performance of participants could be classified through using strong and weak hand shooting scenarios, although the model performance between participants varied widely. However, it was possible to separate the different phases of a shooting scenario through pose estimation with high accuracy. This was done using a Vision Transformer (ViT) model, which could explain its predictions on a per sample basis through the attention mechanism. These explanations were processed and visualised to be interpretable by humans by presenting skeletons in an interactive 3D environment with continuously shifting attention values per joint over time. Although there is much more work that can be done in the sports shooting domain with body tracking data, this study has laid a groundwork for future studies to build on.

## Acknowledgments

We wish to thank the anonymous reviewers, whose suggestions helped improve and clarify this manuscript. Thank you to Max Pettersson and Saga Bergdahl, whose ideas and collaboration in the execution of this study were valuable and rewarding. A big thank you to Anders Johanson, Per Lexander, Olof Bengtsson, Robert Andrén, and everyone else at Saab AB, Training and Simulation who have assisted in the data collection and idea stages of this study. Thank you also to the 13 anonymous participants who took part in the data collection. This work has been performed within the *Mining Actionable Patterns from complex Physical Environments* (MAPPE) research project in collaboration with Jönköping University and Saab AB, Training and Simulation, and was funded by the Knowledge Foundation (research grant number 20180191).

## References

- [1] B. Mason, L. Cowan and T. Gonczol, *Factors Affecting Accuracy in Pistol Shooting*, 1990, pp. 2–6.
- [2] H. Aalto, I. Pyykkö, R. Ilmarinen, E. Kähkönen and J. Starck, Postural stability in shooters, *ORL* **52**(4) (1990), 232–238.
- [3] K. Mononen, N. Kontinen, J. Viitasalo and P. Era, Relationships between postural balance, rifle stability and shooting accuracy among novice rifle shooters, *Scandinavian Journal of Medicine and Science in Sports* **17**(2) (2007), 180–185.
- [4] P. Era, N. Kontinen, P. Mehto, P. Saarela and H. Lyytinen, Postural stability and skilled performance—a study on top-level and naive rifle shooters, *Journal of Biomechanics* **29**(3) (1996), 301–306.
- [5] G. Sattlercker, M. Buchecker, E. Müller and S.J. Lindinger, Postural balance and rifle stability during standing shooting on an indoor gun range without physical stress in different groups of biathletes, *International Journal of Sports Science and Coaching* **9**(1) (2014), 171–183.
- [6] K.A. Ball, R.J. Best and T.V. Wrigley, Body sway, aim point fluctuation and performances in rifle shooters: Inter- and intra-individual analysis, *Journal of Sports Sciences* **21**(7) (2003), 559–566.
- [7] B. Goldberg, C. Amburn, C. Ragusa and D.-W. Chen, Modeling expert behavior in support of an adaptive psychomotor training environment: a marksmanship use case, *International Journal of Artificial Intelligence in Education* **28** (2017), 9s.
- [8] D. James and J. Dyer, Rifle marksmanship diagnostic and training guide, Arlington: U.S. Army Research Institute for the Behavioral and Social Sciences, 2011, p. 123.
- [9] G. Johansson, Visual perception of biological motion and a model for its analysis, *Perception & Psychophysics* **14**(2) (1973), 201–211.
- [10] J. Shotton, T. Sharp, A. Fitzgibbon, A. Blake, M. Cook, A. Kipman, M. Finocchio and R. Moore, Real-time human pose recognition in parts from single depth images, *Communications of the ACM* **56**(1) (2013), 116–124.
- [11] Y. Du, W. Wang and L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June-2015, 2015, pp. 1110–1118.

- [12] Q. Ke, M. Bennamoun, S. An, F. Sohel and F. Boussaid, A new representation of skeleton sequences for 3d action recognition, In *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, 2017, pp. 4570–4579.
- [13] L. Xia, C. Chen and J.K. Aggarwal, View invariant human action recognition using histograms of 3d joints, In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012, pp. 20–27.
- [14] Y. Tang, Y. Tian, J. Lu, P. Li and J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5323–5332.
- [15] S. Yan, Y. Xiong and D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018, pp. 7444–7452.
- [16] K. Otte, B. Kayser, S. Mansow-Model, J. Verrel, F. Paul, A.U. Brandt and T. Schmitz-Hübsch, Accuracy and reliability of the kinect version 2 for clinical measurement of motor function, *PLoS ONE* **11**(11) (2016).
- [17] D. Leightley, M.H. Yap, J. Piasecki, M. Piasecki, J. Cameron, Y. Barnouin, J. Tobias and J. Mcphee, Postural stability during standing balance and sit-to-stand in master athlete runners compared with non-athletic old and young adults, *Journal of Aging and Physical Activity* **25** (2016), 1.
- [18] B. Dehbandi, A. Barachant, A.H. Smeragliuolo, J.D. Long, S.J. Bumanlag, V. He, A. Lampe and D. Putrino, Using data from the microsoft kinect 2 to determine postural stability in healthy subjects: A feasibility trial, *PLoS ONE* **12**(2) (2017).
- [19] R.A. Clark, Y. Pua, C.C. Oliveira, K.J. Bower, S. Thilarajah, R. McGaw, K. Hasanki and B.F. Mentiplay, Reliability and concurrent validity of the microsoft xbox one kinect for assessment of standing balance and postural control, *Gait and Posture* **42**(2) (2015), 210–213.
- [20] A. Elaoud, W. Barhoumi, E. Zagrouba and B. Agrebi, Skeleton-based comparison of throwing motion for handball players, *Journal of Ambient Intelligence and Humanized Computing* **11**(1) (2020), 419–431.
- [21] J. Ko, D. Han and K.M. Newell, Skill level changes the coordination and variability of standing posture and movement in a pistol-aiming task, *Journal of Sports Sciences* **36**(7) (2018), 809–816.
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser and I. Polosukhin, Attention is all you need, In *Advances in Neural Information Processing Systems*, volume 2017-Decem, Neural information processing systems foundation, 2017, pp. 5999–6009.
- [23] L.F. Yeung, K.C. Cheng, C.H. Fong, W.C.C. Lee and K. Tong, Evaluation of the microsoft kinect as a clinical assessment tool of body sway, *Gait and Posture* **40**(4) (2014), 532–538.
- [24] F. Sgrò, S. Nicolosi, R. Schembri, M. Pavone and M. Lipoma, Assessing vertical jump developmental levels in childhood using a low-cost motion capture approach, *Perceptual and Motor Skills* **120**(2) (2015), 642–658.
- [25] Z. Gao, Y. Yu, Y. Zhou and S. Du, Leveraging two kinect sensors for accurate full-body motion capture, *Sensors* **15**(9) (2015), 24297–24317.
- [26] J.C. Núñez, R. Cabido, J.J. Pantrigo, A.S. Montemayor and J.F. Vélez, Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition, *Pattern Recognition* **76** (2018).
- [27] R. Vemulapalli, F. Arrate and R. Chellappa, Human action recognition by representing 3d skeletons as points in a lie group, In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.
- [28] J. Liu, A. Shahroudy, D. Xu and G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9907 LNCS, 2016.
- [29] J. Liu, G. Wang, P. Hu, L.Y. Duan and A.C. Kot, Global context-aware attention LSTM networks for 3D action recognition, In *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, 2017.
- [30] S. Song, C. Lan, J. Xing, W. Zeng and J. Liu, An end-to-end spatio-temporal attention model for human action recognition from skeleton data, In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, 2017.
- [31] C. Plizzari, M. Cannici and M. Matteucci, Spatial temporal transformer network for skeleton-based action recognition, 2020.
- [32] W. Shadish, T. Cook and D. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Houghton Mifflin Harcourt, 2002.
- [33] M. Pettersson, Finding actionable patterns in eye movement and rifle handling, 2021.
- [34] *Saab AB, Training and Simulation*, 5 2019. Live fire training – Ready for any encounter, ver. 1.
- [35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An image is worth  $16 \times 16$  words: Transformers for image recognition at scale, In *International Conference on Learning Representations*, page to appear, 2021.
- [36] J. Sim and C.C. Wright, The kappa statistic in reliability studies: Use, interpretation, and sample size requirements, *Physical Therapy* **85**(3) (2005), 257–268.