Project 2 CS-433 - Efficiency of markless pose estimation

Adrien Feillard (NX), Lou Fourneaux (NX), Aurélien Laissy (GM)
EPFL

***Abstract*—Nowadays, motion capture of individuals is a central theme. Our project supervisor, David Rode, is conducting a Ph.D. on the applicability of markerless technologies for physiotherapeutic applications. One goal is to study the viability of BlazePose to monitor and supervise patients exercising from home. BlazePose, a pose detection model developed by Google, computes 3D coordinates. Current alternatives pose costs and setup problems which makes it hard to implement for home-based physiotherapy.**

This project's primary objective is to determine the accuracy of markerless systems in detecting joints during exercises using machine learning classifying models. Markerless pose estimation involves determining the position and orientation of body joints. It has the potential of providing an affordable and user-friendly solution.

I. Introduction

To assess the effectiveness of the markerless system BlazePose, it is crucial to evaluate how accurately we can predict the type of exercise performed. BlazePose is part of an open-source framework for building cross-platform computer vision and machine learning pipelines. Specifically designed for accurate and real-time pose estimation, it can detect and track key points on the human body in video streams. It provides a solution for various applications, including fitness tracking, gesture recognition, etc. as it is capable of detecting multiple key points representing different parts of the body, such as the eyes, shoulders, elbows, etc. [3].

The first aim of this project is to classify the type of exercise performed by the participants. The second task is the classification of both the exercises performed and the types of errors made while practicing. To achieve this, we developed neural networks and classifier models capable of determining the type of exercises based on 3D joint positions. The different exercises and mistakes are listed in Table 1 below.

Before applying any machine learning techniques, a comprehensive data exploration was conducted. This phase facilitated the identification of pertinent features along with the requisite pre-processing steps for appropriate treatment. Following this initial step, we embarked on evaluating a variety of machine learning models covered in our course. Then we implemented cross-validation as a means to gauge our model's effectiveness and its capacity for generalization to new data.

II. Preparation

A. Dataset

The data set was created by recording 25 unimpaired participants doing 3 sets of 7 different exercises among which

EXERCICES	MISTAKE 1	MISTAKE 2	MISTAKE 3	MISTAKE 4	MISTAKE 5
Squat	Leg axis incorrect	Body facing forward	Back not straight		
Shoulder rotation & elbow extension	Angle between thorax and humerus <90°	Not 90° in elbow	Upper body not straight	Not 90° abduction in shoulder	Not 90° external rotation in shoulder
Knee extension seated	Upper body leaning back	Movement not executed over full motion range			
Supported quadriceps stretch	Knees not next to each other	Arched back	Hand holding foot	Upper body leaning towards supporting leg	Knees not next to each other
Shoulder bridge single leg	Heels not place below knees	Hips not neutral	Knees wider than hips during movement	Feet too close	Stretched out leg higher than supporting leg
Bird dog	Back not neutral	Supporting knee not below hips	Supporting hand not under shoulder	Neck not neutral	
Hip abduction side-lying	Range of motion larger than 45°	Moving leg in front of body	Toes pointing to sky		

Fig. 1: Exercises and mistakes performed by the participants

on is correct and 2 are incorrect. It is composed of about 2.2 million frames captured by 4 webcams. The reference frame is chosen at the waist of the participants. Every frame of the dataset includes the following information:

- The participant's ID
- The type of exercise performed
- Correctness of the exercise: correct or mistake type
- The camera who took this frame
- The time frame in second (every 0.033s)
- The position of 33 joints in the X,Y and Z planes

The recorded joints in Blazepose are the one labelled in figure 2 in blue are the ones present in our dataset. Vicon was used as a comparison to produce the dataset and quantify the position error with Blazepose.

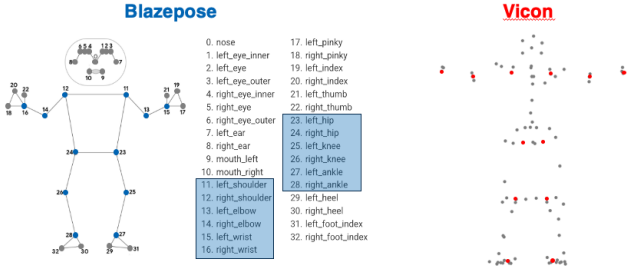

Fig. 2: Comparison of common joints in Blazepose and Vicon (highlighted in blue)

B. Data Analysis and Preprocessing

- *Continuous/categorical features:* The first features are categorical but can be dealt with by encoding them as they represent few possibilities. For example, there are seven exercise types, each of which, can be encoded by a value between 1 and 7.
- *Missing values:* Some input frames were composed of Nan values. They had have to be deleted as they are not relevant.
- *Exploration of Irrelevant Features:* Few features were unrelated to our research objectives: the camera angle, the participant number or the time at which the joints' positions were recorded. Some were kept for the implementation of the GRU model as explained later. For the others models they were suppressed resulting in 99 features.
- *Normalization:* The joints' positions were normalized by Blazepose based on the origin at the waist. Therefore, the coordinates have positive and negative values. No further normalization was needed.
- *Preprocessing of task 1 and 2:* The two tasks of classification only differed in the preprocessing step. For task 1, the feature 'Exercise' was encoded with values ranging from 0 to 7 whereas for task 2, we fused the columns 'Exercise' and 'Set' which deals with the mistakes (between 2 to 5 different mistakes depending on the exercises) and encoded this feature with 38 discrete values. This allowed us to modulate the output size of the different neural networks and the classifier used. The input data was processed accordingly to the their use in the different methods and their specificity are explained in the following section.

III. METHODS

In order to get the most accurate results possible, we tried several methods that could potentially be applied to our problem. After discussing this with our supervisor, we decided to submit every method on a basis of exploratory research. Making sure our predictions were the best possible and not the best ones a single model could produce was crucial in evaluating Blazepose's accuracy. It can exclude the possibility of thinking BlazePose is an inaccurate pose estimation system instead of an ill-suited classification model being the cause of bad results. For ease of use of our models, they were all saved after tuning and training and are at the user's disposal through an interactive run.py file.

There are three different neural networks models and one Random Forest classifier model. All our neural networks were trained using the Cross Entropy Loss criterion and the Adaptive Moment Estimation (Adam) optimization algorithm. The selection of Adam was primarily motivated by its efficient computational performance during development. For the loss criterion, we opted for PyTorch's *nn.CrossEntropyLoss*, which integrates *nn.LogSoftmax* and *nn.NLLLoss* into a single class. This configuration proves effective for optimizing gradients and applying the softmax activation function to the output layers of neural networks. The chosen setup is particularly well-suited for addressing multiclass classification problems.

For each of the methods and tasks, a different processing of the data was applied since the inputs needed were different. Afterwards, we proceeded to tune the hyperparameters specifically to each method using grid search cross validation on a validation set consisting in 15% of the original dataset. Outside of the specific parameters for each method, different number of neurons for each neural networks, as well as the activation function of the layers. An early stopping callback was also implemented to decrease computational cost during tuning. Finally, they were trained on 70% of the original dataset and tested on 15% of the original dataset.

A. *Neural Networks (NN)*

We first tried to implement a simple fully connected neural network as they usually work well for input data that has a relatively simple structure.

- **Model architecture**: The class is composed of 3 layers where each layer is composed of a different number of neurons. The number of neurons for each layer and the activation function was determined trough tuning while the number of layers and the leaning rate were determined though experimentation. In fact, the tuned paramteres vary depending on the classification task being performed (1 or 2).
- **Specific feature processing**: In this model, only the the joints' coordinates were kept for the input data resulting in 99 input features.

	Epochs	lr	N1	N2	Activation	Batch	k-Folds
Task 1	100	0.001	2048	384	Sigmoid	151	3
Task 2	100	0.001	4096	512	ReLU	151	3

TABLE I: Tuned hyperparameters for the NN model for each task

However, taking into account the sequential aspect of our data, we thought of a recurrent neural network. Indeed, fully connected neural networks could be limited when dealing with sequential data or spatial information which is the case of our data set (article [2] on deep learning). So even though this network has a great accuracy, we explored the possibility of a special type of recurrent neural network.

B. *Gated recurrent unit (GRU)*

A GRU is a recurrent neural network which is often well-suited for sequential data such as time series because it captures the sequential dependencies. We used this model because the frames were timed for each particpants and exercise. GRU model needs at least one feature linearly dependent for each sequences.

- **Model architecture**: The model is composed of only two layers where the first one is GRU and the second one is linear. The activation function can be either Sigmoid, ReLU or Tanh determined by tuning along with the number of neurons for the layer 1.

- **Specific feature processing**: In the data pre-processing, encoded camera angle and participants as well as time were kept to ensure that the model could recognize the different sequences in the the dataset. It constitutes the only model where the dataset was not shuffled.

Epochs	lr	N1	Activation	Batch
100	0.001	64	ReLU	151

TABLE II: Default parameters for the GRU model for both tasks

This method is very high in terms of computational costs, therefore, we only trained the model based of default parameters instead of tuned ones. Its computational expense is due to comparison of inputs with previous data to establish connections. Because, it didn't yield a substantial improvement in performance compared to alternative models and because no shuffling could be done (the model could be trained on only one part of the participant's exercise) which could lead to underfitting, we decided to explore alternative network architectures.

C. *Convolutional Neural Networks (CNN)*

- **Model architecture**: The model comprises two 3D convolutional layers for spatial relationships, a max pooling layer for dimension reduction, and two fully connected linear layers for feature integration and transformation, using the same activation function throughout, with the final layer handling multi-class output.
- **Specific feature processing**: Utilizing a $3 \times 3 \times 11$ grid map, features are placed in cells based on their mean 3D position, facilitating the analysis of relationships between close features while avoiding unnecessary comparisons between distant ones.
- **Variables parametrization**: The model's performance can be adjusted by modifying parameters such as grid map size, kernel filter size, stride, padding, and activation function. A $(3 \times 3 \times 11)$ grid map size was chosen for computational efficiency, balancing detailed performance and computational cost.

	lr	N1	N2	N3	Kernel1	Kernel2	Activation	Batch
1	0.001	64	128	50	(2,2,2)	(3,3,3)	Leaky ReLU	64
2	0.001	64	128	50	(1,1,1)	(3,3,3)	Leaky ReLU	64

TABLE III: Tuned hyperparameters for the CNN model for each tasks

We opted for a Convolutional Neural Network (CNN) for our 3D joint coordinate analysis due to its efficiency in processing grid-like data structures. The CNN's translation invariance is key for consistent joint pattern recognition, regardless of position changes. Its ability to learn hierarchical features is also crucial for understanding complex spatial relationships in the data, making it ideal for analyzing intricate patterns in our dataset.

D. *Random Forest Classifier (RFC)*

After facing computational difficulties, we thought classifiers could be a better fit for our problem. Random Forests are effective for both small and large datasets and are generally robust to noisy data. The parallelization of tree construction makes Random Forest scalable, allowing it to handle large amounts of data efficiently. The feature processing was the same as for the NN model and the variable we parameterized are in the table below. A higher number of trees generally leads to better performance but increases computation time as well as potential overfitting.

Decision trees	Max depth	Random state	Folds
200	None	100	4

TABLE IV: Tuned hyperparameters for the RFC model for both tasks

IV. RESULTS

A. *Exercise classification*

In order to have a more robust dataset, we kept all the sets of exercises whether they were correct or had an error 'A', 'B', etc. This way no feature augmentation was needed and all the benefits were present:

- Increased dataset size: can lead to better model generalization.
- Robustness to variability: the model becomes more robust to variations, such as changes in orientations.
- Reduced overfitting: augmentation acts as a form of regularization by preventing the model from memorizing the specific instances in the training data.

Table V summarizes the accuracies for each model in the case of the first classification task. Accuracy was chosen as the primary metric instead of the F1 score to consider both true positives and true negatives. This decision aligns with the broader objective of evaluating comprehensive accuracy across all classes rather than emphasizing precision and recall alone.

From these results, we can conclude that in recognizing basic movements, BlazePose is an efficient system. Indeed, the exercises were rightly (close to 100%) predicted from the BlazePose acquired data.

Model	NN	GRU	CNN	RF
Training accuracy (%)	99.31	96.38	98.18	99.36
Testing accuracy (%)	98.82	96.20	97.82	99.51

TABLE V: Accuracy of test and training for each model

B. *Exercise and mistake classification*

This task is more interesting to determine the preciseness of the BlazePose system. The models each had pros and cons but the important information is the overall accuracy across the different classification techniques.

Random Forest Classifier is the best performing model for task 1 and 2. However it is also the one that has the highest computational cost with respect to the other models. GRU is the least accurate model, as it is the worst in both cases. The Neural Network and Convolutional Neural Network are good alternatives for the first and second tasks.

Although for task 1, all methods have fairly good performance, results for task 2 are more disparate. The differences

in accuracy are greater between the models and between training and testing phases. Specifically, there is an observed maximum difference of 3.04%, suggesting the possibility of overfitting in certain models. Nonetheless, this difference did not appear substantial enough to introduce regularization, especially considering the absence of overfitting in task 1. This could be improved by feature augmentation.

Model	NN	GRU	CNN	RF
Training accuracy (%)	92.69	55.20	77.01	95.75
Testing accuracy (%)	89.65	54.86	74.85	96.71

TABLE VI: Accuracy of test and training for each model

Tracking the accuracy evolution of the models proves to be an insightful visualization, helping to discern whether enhancements in model performance can lead to improved predictions or whether the peak accuracy attained is constrained by the limitations inherent in the capture of BlazePose's joint coordinates (fig 3). For example, task 1 seems to have reached a maximum accuracy whereas for task 2 no plateau has been reached except for GRU so the models could be further improved.

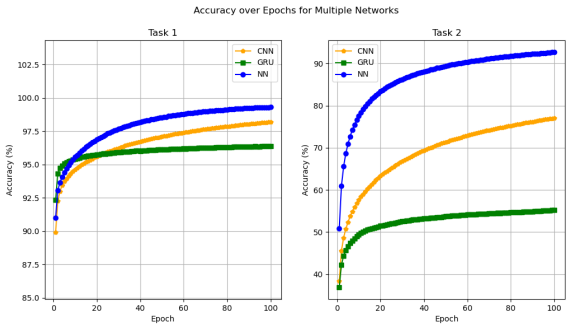

Fig. 3: Accuracy percentage evolution for each epochs of the NN, GRU and CNN neural networks

Confusion matrices can tell us more about which exercises were especially recognizable using the estimation system and in which exercises some problems might occur (fig 4).

Confusion matrices pinpoint model weaknesses in movement identification, with the supported quadriceps stretch (indices 13 to 18) posing the greatest challenge. Errors are difficult to discern, especially in detecting nuanced actions like arched back, where limited monitored joints, particularly in the back, contribute to detection challenges.

V. Discussion

Task 2 exhibited lower prediction accuracy than Task 1, possibly due to BlazePose's limitations in recognizing complex movements and its inherent challenges in accurately detecting joints. The similarity of movements and the potential difficulty in distinguishing mistakes from correct execution, exacerbated by BlazePose's limited accuracy, could contribute to model errors. Additionally, the increased complexity with 39 classes in Task 2 may further complicate decision-making for the

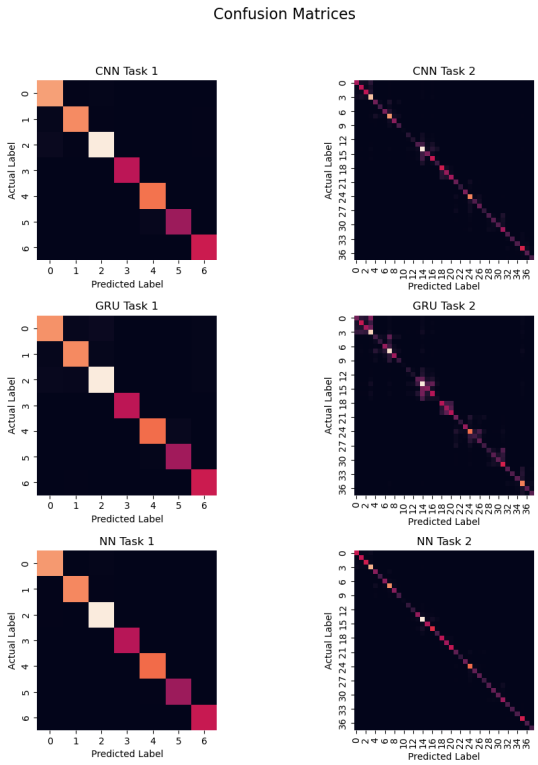

Fig. 4: Confusion matrices of the different method for all the tasks

models, contributing to the observed decline in accuracy. Considering these factors, improving accuracy in Task 2 remains a challenge.

VI. Summary

The project aimed to evaluate BlazePose's efficacy in markerless pose estimation. In the initial task, the system demonstrated exceptional performance, achieving near-perfect accuracies of almost 100% and showcasing its proficiency in recognizing basic movements. In the subsequent task involving more intricate exercises, BlazePose consistently proved effective across various models. However, it exhibited a slightly diminished performance compared to Task 1 when dealing with similar movements.

Reflecting on the project's overall success, the high accuracy in Task 1 underscores BlazePose's efficiency in recognizing fundamental movements. Task 2's challenges, on the other hand, highlight potential limitations in handling more complex exercises. Future work could concentrate on refining BlazePose's pose estimation accuracy, exploring advanced post-processing techniques, and investigating alternative models to address specific challenges observed in the project. Despite these challenges, the project provides valuable insights into the capabilities and limitations of BlazePose in markerless pose estimation, paving the way for advancements in home-based physiotherapy applications.

ANNEX

ETHICAL RISKS

In our pursuit of developing a machine learning model for accurately assessing physiotherapy exercises using joint coordinates, we conducted an Ethical Risks Assessment using the Canva framework, revealing several ethical concerns. The foremost among these is the potential bias towards a specific demographic group, namely children, categorized as a fairness risk.

The dataset exclusively comprises healthy, unimpaired young adults with an average age of 26, deliberately excluding measurements from children. Consequently, there is a risk of less accurate predictions for kids, which could lead to misclassifications and the possibility of exercises being deemed correct/incorrect even when performed correctly/incorrectly. Such inaccuracies could instill poor movement habits in children, exacerbating the quality of their physiotherapy treatment and, subsequently, impacting their health.

To quantify fairness, we aimed to compute the Demographic Parity Difference, but it was hindered by missing sex and age information, a critical barrier. However, a validation study led by David Rode with impaired patients is underway, contributing to the representativeness of elderly individuals in the dataset, although it doesn't fully address the risk for children. With a 13% height difference between a 12-year-old Swiss boy (estimated at 150 cm [4]) and a 26-year-old (male or female) at 172.5 cm [5], a potential mitigation strategy involves training the model with a normalized dataset. This entails translating joint coordinates proportionally by a factor 0.87 towards the hip, the coordinate center, to create synthetic data representing children and improve the model's handling of their unique features. Comparing results with a model trained with out data set we observed a very slight accuracy deacrease for NN on task 2 89.64% with respect to 89.65%, thus clearing the importance this risk. However, this approach assumes that children share the same flexibility and proportions as adults. However this approach is criticized as it was used to model women in car crash using the reference of male bodies and led to erroneous results. [1] The best solution would be to use children to create the dataset.

In conclusion, mitigating these ethical risk involves efforts to enhance dataset representativeness physiotherapy exercises assessment.

On figure 5, there is no difference with NN on task 2, without the normalization of 0.87. Thus this risk is minimal.

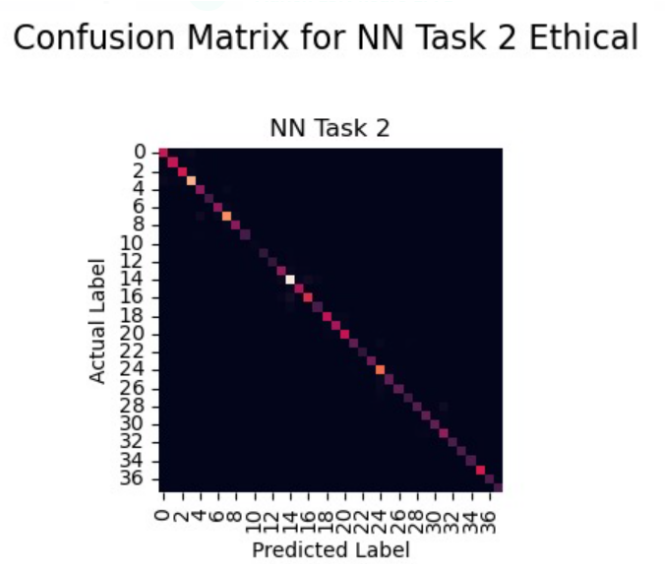

Fig. 5: Confusion Matrix of a 0.87 Normalized test set

REFERENCES

[1] Inclusive crash test dummies: Rethinking standards and reference models.

[2] Zhang J. Humaidi A.J. et al. Alzubaidi, L. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J Big Data 8*, 2021.

[3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Lixuan Zhu, Fan Zhang, and Matthias Grundmann. Blazepose: On-device real-time body pose tracking. *ArXiv*, abs/2006.10204, 2020.

[4] Hélène Kaufmann, Albert Rieben, and Richard Lang. *Croissance de la taille et du poids de 4 à 19 1/2 ans: garçons et filles suisses domiciliés dans le canton de Genève en 1972*. Médecine et Hygiène, Genève, 1976.

[5] Office fédéral de la statistique (OFS). Taille des jeunes adultes suisses, 2022.

Fig. 6: Ethical Risks Assessment Canva

Digital Ethics Canvas 2023 - Data Science Version - C. Hardebolle

Dataset

□ Who created the dataset?

David Rode, a PhD student at SMS lab of ETH.
David's research focuses on Motion Capturing and Pose Estimation with markerless systems and their use in therapeutic settings.

□ For what purpose was the dataset created?

There was not enough data available for his research, so David Rode had to create some in order to determine accuracy of markerless systems at detecting joints, measuring joints angles,identifying type of movement, and at rating exercise quality when performing exercises.

□ What mechanisms or procedures were used to collect the data?

The new dataset of poses during exercises consists of 25 unimpaired participants doing 3 sets of 7 different exercises, among which 1 is correct and 2 are incorrect.

The experiment setup was 27 Vicon cameras (ground truth system), and 4 webcams (for the Blazepose system) recording the participants during exercises.

□ Who was involved in the data collection process?

David Rode, a physical therapist and a movement specialist.
Akina, a Zurich based medtech startup and ETH Zurich sponsored the study.

□ Over what timeframe was the data collected?

Between May 2023 and November 2023.

□ Was any preprocessing of the data done?

The normalization is done by Blazepose. Blazepose has two possible output formats: global and image coordinates.
In case of image coordinates, the distances are output normalized to the image size. The depth information is normalized with respect to the location of the hip centerpoint.
No further preprocessing was done on the dataset.

□ Are there any missing data or data errors?

Yes, they were removed during preprocessing, but a negligible amount.

□ Where is the data stored?

Data is stored locally on David Rode's device, the ETH server with secure access and on the ETH sharepoint he shared with our team.

Beneficence

□ What are the expected benefits of analyzing this data? For whom?

Estimate the accuracy of the Blazerpose thus help decide if markerless systems are viable for home therapy. Enable patients to do therapy from home. Will help reducing the costs of treatment for patients since they will do the exercises using a software rather than going to the physiotherapist.

Privacy

Risks	Mitigation

□ Does the data contain personal or sensitive information?

The data we received is already processed so it is not a video anymore. It is sensitive as it is biometric data of joint coordinates of people doing exercises. Our dataset is anonyme but the data that will be collected by the system won't be anonyme as it would provide personal feedback like a physiotherapist would.

□ Can personal or sensitive information be derived or inferred from the data or from the analysis?

Yes we can derive personal information. The dataset contains people's dimensions (positions of joints in 3D space). We could deduce the sizes of body parts. This is sensitive data as this data could be used to send targeted adds (one could imagine advertising long clothes if the person was identified as tall for instance). From the results we can also derive if the person executes well exercises or deduce the types of exercises done thus gaining insights about the health of the patient which is a really sensitive data. For example we could identify tremors and infer that person has Parkinson. Which is a valuable information for an insurance company. Thus is dataset if not anonymized can be really sensitive and the privacy risk is high.

Non-maleficence

Risks	Mitigation

□ Does the dataset contain unsafe data (violence, nudity…)?

No, just the coordinates of the joint's position from the participants.

□ What kind of impacts can errors in the data or in the analysis have?

Missing data can lead to less precision in the system which in turn can lead to bad identification of the movement or overlooking incorrect executions. The patient could injure himself, or develop a muscular disequilibrium despite receiving good feedback from the system.

□ Could the data or the conclusions from the analysis be used in harmful ways?

No, it will just help decide if the markerless pose estimation is viable or not.

Fairness

Risks	Mitigation

□ Is the data representative from a larger set (population)? How are subgroups represented?

The dataset only represents healthy individuals. Age distribution was biased towards younger participants due to recruitment among peers.
There are 11 Males and 14 Females
The average age is 26, and features only young adults (children were excluded, max age was 33) This means that the data is not representative of impaired people. The kids and elderly people are also not represented.

□ What kinds of biases may affect the data?

Since age distribution is biased towards young adults, it means that children, impaired and elderly peoples behaviors weren't monitored. So they will be more prone to have a worse accuracy than young adults due to their different posture and size. As a matter of facts young adults are taller than children and tend to have a more upright posture than elderly people. The data is biased in terms of the height of the participants.

□ Can the outcomes of the analysis be different for different groups?

Yes given that there is a bias for some groups it is very likely that the outcome will be different for them. As a matter of fact some exercises could be misclassified or the exercise could be done correctly/incorrectly and still be classified as correct/incorrect.

□ Could the data or analysis results contribute to discrimination against people or groups?

This would not contribute to discrimination, nevertheless some groups could have a better/worse experience with Blazepose. This would involve a bad physiotherapy treatment and could worsen their health.

Sustainability

Risks	Mitigation

□ What is the carbon and water footprint generated by the storage of the data and by the computation in the analysis process?

It sure consumes energy and produce CO2 to store this data. But the dataset is still reasonably small. The computation process is pretty straightforward once training of the models have been done.

□ What type of human manual labor is involved in the data (e.g. labeling)?

Planning the movements to be executed by the participants and labeling the exercises.

□ Does the data or the analysis require updates?

To be able to recognize more exercises or account other types of incorrect execution we could have to train the data with those new features to ensure we can classify them. Occasionally the dataset will also be updated to fix errors in postprocessing Vicon or different settings for the markerless pose estimator.

Empowerment

Risks	Mitigation

□ How are the people concerned involved with the data or the analysis: have they been notified, have they consented?

All subjects were made aware of their rights according to ETH ethics regulations. They signed a form and accepted that their motion data will be used for research and development. Image rights were requested too and only images of subjects who accepted this, were shared.

□ Are the people concerned able to make choices (e.g. revoke consent, modify or delete data) regarding the data or the analysis?

Yes, the study was conducted according to ETH ethics guidelines, which conform to the regulations in Switzerland.
Participants can request viewing their personal data but cannot publish these without approval. Consent was given at time of the measurement. At any time during the measurement consent could be revoked. After capture, data can be used for research and development and may be published, if no conclusions on the identity of the participants can be made. Deletion of data retrospectively was not among the listed rights.