



Fast and robust video-based exercise classification via body pose tracking and scalable multivariate time series classifiers

Ashish Singh¹ · Antonio Bevilacqua¹ · Thach Le Nguyen¹ · Feiyan Hu² · Kevin McGuinness² · Martin O'Reilly³ · Darragh Whelan³ · Brian Caulfield⁴ · Georgiana Ifrim¹

Received: 10 November 2021 / Accepted: 11 November 2022 / Published online: 21 December 2022
© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2022

Abstract

Recent technological advancements have spurred the usage of machine learning based applications in sports science and healthcare. Using wearable sensors and video cameras to analyze and improve the performance of athletes, has become widely popular. Physiotherapists, sports coaches and athletes actively look to incorporate the latest technologies in order to further improve performance and avoid injuries. While wearable sensors are very popular, their use is hindered by constraints on battery power and sensor calibration, especially for use cases which require multiple sensors to be placed on the body. Hence, there is renewed interest in video-based data capture and analysis for sports science. In this paper, we present the application of classifying strength and conditioning exercises using video. We focus on the popular Military Press exercise, where the execution is captured with a video-camera using a mobile device, such as a mobile phone, and the goal is to classify the execution into different types. Since video recordings need a lot of storage and computation, this use case requires data reduction, while preserving the classification accuracy and enabling fast prediction. To this end, we propose an approach named BodyMTS to turn video into time series by employing body pose tracking, followed by training and prediction using multivariate time series classifiers. We analyze the accuracy and robustness of BodyMTS and show that it is robust to different types of noise caused by either video quality or pose estimation factors. We compare BodyMTS to state-of-the-art deep learning methods which classify human activity directly from videos and show that BodyMTS achieves similar accuracy, but with reduced running time and model engineering effort. Finally, we discuss some of the practical aspects of employing BodyMTS in this application

Responsible editor: Indre Zliobaite.

✉ Ashish Singh
ashish.singh@insight-centre.org

Extended author information available on the last page of the article

in terms of accuracy and robustness under reduced data quality and size. We show that BodyMTS achieves an average accuracy of 87%, which is significantly higher than the accuracy of human domain experts.

Keywords Video-based exercise classification · Strength and conditioning · Body pose tracking · Time series classification

1 Introduction

Recent years have seen a tremendous growth of the use of machine learning for sports science and healthcare applications. This is mainly due to the increased usage of wearable sensors and video-based tracking devices (Ahmadi et al. 2014; O'Reilly et al. 2015, 2017, 2018; Fawaz et al. 2019; Kwon et al. 2020; Choutas et al. 2018) to capture data that is utilized for rehabilitation or to assess the performance of athletes (Richter et al. 2021).

Human exercise performance classification is a sub-field of human activity recognition (HAR) where the goal is to classify the execution of an exercise into predetermined classes. Most research in this field has focused on utilizing inertial sensors for data capture (Ahmadi et al. 2014; O'Reilly et al. 2015, 2017, 2018; Fawaz et al. 2019), which commonly involves extracting domain-specific or predefined statistical features from sensor data and applying supervised machine learning methods. However, using sensors to collect human activity data has some notable limitations: sensor-based data collection is error-prone and time-consuming as sensors require careful positioning on the body, as well as calibration for the specific task (Whelan et al. 2016; Kwon et al. 2020).

This work focuses on classifying physical exercise execution by using video data capture. Video data helps to alleviate some of the above problems, as videos can be easily captured through available smartphones and data capture does not require multiple specialized sensor devices to be worn on the body, thus avoiding issues such as discomfort and impeding the ease of movement (Kwon et al. 2020). In this paper, we work with video recordings of participants executing the Military Press (MP) exercise. MP is an important exercise in strength and conditioning, injury risk screening, and rehabilitation (Whelan et al. 2016). The main objective is to classify exercise performance in terms of differentiating between correct and different aberrant executions of the exercise. Incorrect execution may lead to musculoskeletal injuries and impede performance (Baechle and Earle 2008), therefore, automated and accurate feedback on execution is important to avoid injuries and maximize the performance of the user. While this is an important exercise, it is also a difficult one to classify, with human inter-rater agreement at about 60% (Whelan et al. 2019).

Our previous work (Singh et al. 2020) proposed an approach for interpretable classification of Military Press exercises using videos as time series. We showed that a body pose estimation method, OpenPose (Cao et al. 2019), combined with multivariate time series classifiers (MTSC) can be used to accurately classify and interpret correct and incorrect executions. We henceforth name this approach BodyMTS (for Body tracking Multivariate Time Series). Figure 1 shows the overall flow of BodyMTS: (1)

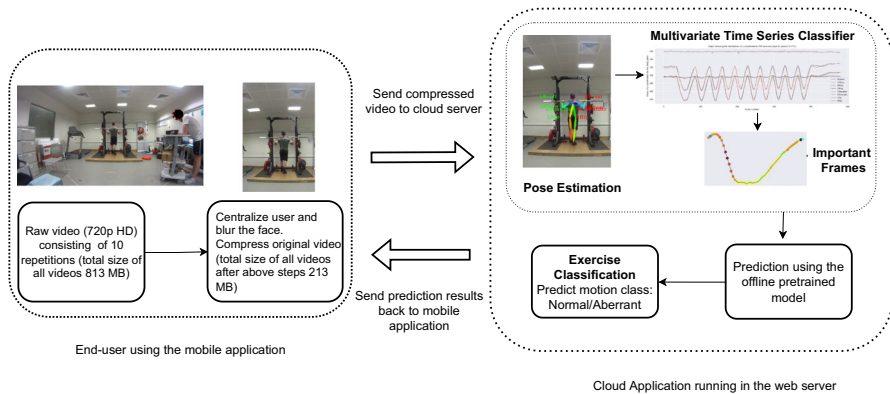


Fig. 1 Overview of the BodyMTS approach for the Military Press strength and conditioning exercise classification. Left-to-right flow: raw video, extracting and tracking body parts using human pose estimation, preparing the resulting data for multivariate time series classification and interpretation

pose estimation identifies and tracks multiple body parts over the video frames, (2) the (X, Y) location coordinates of body parts for each frame are extracted resulting in multivariate time series, (3) a multivariate time series classifier is trained to classify the execution of the exercise into pre-defined classes.

In this paper, we extend our prior work with an extensive analysis of the robustness of BodyMTS to different sources of data noise, as well as a side-by-side comparison with state-of-the-art deep learning methods for human activity classification directly from videos. Our hypothesis is that body pose estimation provides a strong prior for the classifier which now focuses on the pose information important for the task, not on other details in the video, e.g., background. This is contrasted to direct end-to-end video classification with deep learning, where noisy data may affect the model robustness and accuracy, and generalisation beyond benchmarks is known to be a challenge¹ (Azulay and Weiss 2019). In our experiments, we show that deep learning models require pre-training on large amounts of data, with a gap of 60% in accuracy between training from scratch and pre-trained models.

Although BodyMTS in its current form is a proof-of-concept, we demonstrate its applicability and feasibility by considering the key factors that may influence performance, such as the impact of realistic noise types on the classifier accuracy and running time, as well as the computational resources and storage space used by the data and models. We focus our attention on noise coming from changes in video quality, pose estimation quality, or time series data pre-processing.

While research on assessing the performance of athletes using sensors has been successfully deployed,² there are currently not many approaches to classify the execution of strength and conditioning exercises using videos. In our search, we have identified software such as Kinovea (Adnan et al. 2018; Puig-Diví et al. 2019) and DartFish (Fathallah Elalem 2016; Faro and Rui 2016), which seem to work through manual

¹ <https://www.nature.com/articles/d41586-019-03013-5>.

² Two of the co-authors of this paper have successfully launched Output Sports, a start-up built on commercialising research results on single sensor systems: <https://www.outputsports.com>.

analysis at a very low frame rate. Despite providing a vast number of features, these systems are not equipped with automatic classification of physical exercises (Adnan et al. 2018; Puig-Diví et al. 2019).

Existing research on human activity recognition from videos is based on applying complex deep learning architectures (Ji et al. 2010; Simonyan and Zisserman 2014; Tran et al. 2015; Feichtenhofer et al. 2019). Despite the competitive performance on benchmarks, this is achieved at the cost of heavy computation resources such as several hours of training and testing on high-end GPU hardware. Besides the need for high-end hardware, this also has a negative environmental effect. Furthermore, these models are trained and tested on datasets such as UCF-101 (Soomro et al. 2012) and Kinetics-400 (Kay et al. 2017), which contain long duration videos and a wide range of activities. For instance, in Kinetics-400 the average duration of a clip is 10 s and the number of samples is around 300k. In our setting, a single clip is of 3 s duration on average and the differences between the classes are subtle, making the classification task more challenging, e.g., cycling versus walking in contrast to executing the MP with/without an arch in the back. Our dataset is also small (a few thousand samples for training and validation) when compared to these large benchmarks. We have found no prior work that uses videos for strength and conditioning exercise classification and works with this type of smaller data scale and fine-grained classification.

Our main contributions in this paper can be summarized as follows:

- We present and extensively evaluate BodyMTS, an end-to-end video-as-timeseries human exercise performance classification method. We study the impact of improvements in body pose estimation methods (e.g., OpenPose, (Cao et al. 2019)) and recent multivariate time series classifiers (e.g., ROCKET Dempster et al. (2020) and MiniROCKET (Dempster et al. 2021)) on the overall classification accuracy. We show improvements in accuracy with an average of 87% classification accuracy for the Military Press exercise.
- We analyze the robustness of BodyMTS against different types of realistic noise and measure the impact on the classifier performance. We consider three common sources of noise in our application setting: video capture quality, pose estimation quality and time series pre-processing steps.
- We conduct an extensive empirical study comparing BodyMTS to state-of-the-art deep learning approaches for human activity recognition directly from videos. We compare all methods in terms of accuracy, training/testing time and computation resources. We show that BodyMTS is robust to lower quality data captured at prediction time and has fast training and prediction.
- To support our paper, all of our code, data and detailed results are available at: https://github.com/mlgig/BodyMTS_2021.git.

The paper is organized as follows. In Sect. 2 we discuss the application and technical requirements of BodyMTS. In Sect. 3, we give an overview of the related literature on human activity recognition, human pose estimation, strength and conditioning exercises and multivariate time series classification. Section 4 describes the data collection process and the Military Press dataset. Section 5, presents our methodology for classifying MP exercises from videos and Sect. 6 describes the main data mining challenges. In Sect. 7, we analyze the robustness of BodyMTS against different sources of noise

and compare its performance with state-of-the-art deep learning methods. In Sect. 8, we describe the lessons learned from this study, as well as limitations and future work. In Sect. 9 we summarise our recommendations for practitioners working on similar tasks and we conclude in Sect. 10.

2 Application requirements

In this section, we discuss the required BodyMTS features and the corresponding application and technical requirements. We note that BodyMTS is currently a proof-of-concept and the actual deployment scenario and requirements may change depending upon the business case and the end-user requirements.

The aim of BodyMTS is to provide a scalable system that can accurately measure and evaluate end-user performance of strength and conditioning (S&C) exercises, with a view to provide feedback in near real-time. This, in turn can guide physiotherapists, trainers, and elite and recreational athletes to perform exercises correctly and therefore minimise injury risk and enhance performance. We devised the following list of application requirements based on previous research in which we consulted with end users, clinicians, and strength and conditioning experts on the design, implementation and evaluation of interactive feedback systems for exercise (Brennan et al. 2020; Argent et al. 2019, 2018; Giggins and Caulfield 2015; O'Reilly et al. 2017):

- Be able to accurately monitor the body parts movement, accounting for the critical body segments involved in the exercise in question.
- Detect when deviations from normal movement profile have occurred, and which kind of deviation has occurred in each case.
- Provide clear and simple feedback to the end user, in near real-time.
- Simple data capture based on ubiquitous sensor technology (e.g. single phone).
- Coverage of wide range of S&C or rehabilitation exercises.

Table 1 summarizes the application features and the corresponding application and technical requirements for such a system.

There are two main components of BodyMTS:

- *Client side mobile application* which the end-user uses to record their execution. The recorded videos are then pre-processed before sending them to the server side running on the cloud.
- *Server side application* which stores the pre-trained model. Each repetition of the clip is classified separately using the stored model. The final results are then returned back to the client side mobile application.

Figure 1 shows the overall workflow of BodyMTS. Before the user starts using the client side mobile component of BodyMTS, we expect the following requirements to be fulfilled:

- The mobile camera used for recording the execution should be placed on a static surface before the start of the execution.

Table 1 Application features, application requirements and associated technical requirements

Application features	Application requirements	Associated technical requirements
Accurate real-time measurement of body part movements involved in the exercise	Level of accuracy and reliability for tracking anatomical landmarks or estimation of joint angles that will facilitate detection of clinically relevant changes in performance of specific exercises. This will differ across exercises and contexts (e.g. greater level of sensitivity required in elite athletes compared to recreational athletes) so an overall benchmark cannot be specified	Accurate real-time human pose estimation for the body parts relevant to the specific exercise; OpenPose (Cao et al. 2019) can detect 25 body parts in an input image in under 1 second and has accuracy above 75% on established benchmarks
System can detect when deviations from perfect movement profile have occurred, and which kind of deviation has occurred in each case	Previous research Whelan et al. (2019) has demonstrated poor intra and inter-rater reliability for domain expert identification of movement deviations in commonly performed screening exercises. Kappa scores at an inter-rater level ranged in 0.18–0.53, and intra-rater agreement ranged in 0.38–0.62. Therefore, establishing a threshold of 0.8 classification accuracy would be a reasonable benchmark for success in this system (i.e., 8 out of every 10 executions are correctly classified)	Accurate and efficient classification: ROCKET (Dempster et al. 2019a; Dhariyal et al. 2020; Pasos-Ruiz et al. 2020) has state-of-the-art accuracy on established MTSC benchmarks and is a very efficient classifier. In our experiments it achieves a classification accuracy higher than 80%
Clear and understandable feedback to end user at an appropriate time (near real-time)	The end-user must receive feedback on the performance of a set of exercise repetitions immediately after the set has been completed. This will require near real-time capture, processing and analysis of the data. The feedback should be meaningful: it should give an explanation as to which part of the movement, and which body locations, are responsible for the observed deviation; should enable the end-user to address this deviation in the subsequent performance of the exercise	Class label feedback: Near real-time class prediction on single video clip. BodyMTS completes data processing and prediction in 25 s for a clip of 10 exercise repetitions. Simple and accurate explanation: Explanation is beyond the scope of this paper

Table 1 continued

Application features	Application requirements	Associated technical requirements
Simple data capture, with ubiquitous sensor technology	Ideally, the system should be based on sensor technology that is very inexpensive, or at least is likely to be already owned by the end user. The system should be easily deployable, and not require extensive setup and calibration prior to each use	Video data capture through mobile phone. Data reduction to enable fast processing, while maintaining accuracy: BodyMTS can reduce the data size by 70% and still preserve classification accuracy higher than 80%
Coverage for wide range of S&C or rehabilitation exercises	The system should be scalable and transferable, in the sense that coverage for additional exercises does not involve an exhaustive data/feature engineering and model training process	Fully automated pipeline: BodyMTS is fully automated. Inclusion of other exercises: This is beyond the scope of this paper

- The view of the camera will vary depending upon the type of exercise (e.g., front view for Military Press). This is static information that will be already stored within the application for each type of exercise.
- The mobile application will use a bounding box to centralize the user with respect to each frame.

These are quality control requirements that are evaluated before and during the execution. The video and pose estimation have to be of sufficient quality, otherwise the data will be rejected by the application. After the above conditions are fulfilled, the user will activate the application and start recording the video using the mobile application. At the end of the workout, the client side application will pre-process the data. The recorded video will be pre-processed to centralize the participant and to remove the audio. Further, it will be compressed to reduce the total size and will be sent to the server side application where further processing takes place. At the server side, the compressed videos undergo pose estimation followed by segmentation and classification steps. Finally, the classification results are returned back to the client side mobile application.

3 Related work

In this section we present an overview of existing approaches for strength and conditioning exercise classification, human action recognition from videos, human pose estimation and multivariate time series classification.

3.1 Strength and conditioning exercise classification

The purpose of S&C exercises is to improve the performance of athletes in terms of strength, speed, flexibility, agility (Trejo and Yuan 2018; O'Reilly et al. 2015, 2018; Chu et al. 2019). S&C exercises span multiple types of exercises or movement sequences that target different parts of the body and different functional goals. In some cases, the person interacts with a weight or mechanical apparatus, whereas in others the person performs a free body movement without any interaction with an external system or force (e.g., jump). Recent advances in technology have spurred the usage of high tech solutions to maximize the performance of athletes. These can be divided into three broad categories: optical motion capture, wearable inertial sensors and video (Singh et al. 2020; Puig-Diví et al. 2019; Faro and Rui 2016; Fathallah Elalem 2016).

The most popular optical motion capture system is Microsoft Kinect. The work of (Trejo and Yuan 2018; Zerpa et al. 2015; Ressler et al. 2020; Decroos et al. 2018; Dajime et al. 2020) has investigated the use of Microsoft Kinect for rehabilitation exercises, movement quality assessment and gait analysis. However, despite their high performance, these systems are expensive, need high maintenance, require significant time to set up and are mostly limited to controlled clinical trials.

Wearable inertial sensors-based approaches consist of fitting Inertial Measurement Units (IMU) (O'Reilly et al. 2018; Chu et al. 2019; Espinosa et al. 2015) on different parts of the body. The sensor data is analyzed to evaluate performance using supervised

machine learning methods, visualization or manual techniques. The number of inertial sensors required and their positions vary from exercise to exercise (Espinosa et al. 2015; Whelan et al. 2016; O'Reilly et al. 2017, 2018). Research methods and also commercial systems have been deployed using such inertial sensors. Still, sensors can be expensive, they may hinder the ease of movement particularly when applied over many body parts and over longer periods of time, and the annotation process can be time-consuming (Whelan et al. 2016; Kwon et al. 2020; Dajime et al. 2020).

The third category uses video-based devices such as dedicated cameras (DSLR) or smartphone cameras to capture data. Proprietary software such as Dartfish (Fathallah Elalem 2016; Faro and Rui 2016) and open-source software such as Kinovea (Adnan et al. 2018; Puig-Diví et al. 2019; Moral-Muñoz et al. 2015) are used to analyze performance by providing the option of slow-motion replay at a very low frame rate. However, these systems are less accurate and require fitting body markers of different color to the background. Recent work (Slembrouck et al. 2020; Nakano et al. 2020a; Stamm and Heimann-Steinert 2020) utilizing pose estimation for motion tracking has paved the way for alternative approaches to IMUs and optical motion capture systems. We found no prior work that utilizes video to classify S&C exercises.

3.2 Human activity recognition

Video-based Human Activity Recognition (HAR) is a core area of computer vision. HAR methods can be broadly classified into two categories. First are methods based on handcrafted features such as bags of visual words (Wang and Schmid 2013; Dalal et al. 2006; Peng et al. 2014; Sánchez et al. 2013). These include finding local spatio-temporal features such as motion boundary histograms (Dalal et al. 2006) and trajectories (Wang and Schmid 2013), followed by feeding them to a classifier. These methods have been shown to provide competitive performance on benchmark datasets (Carreira et al. 2018; Sigurdsson et al. 2016; Soomro et al. 2012) before the emergence of deep learning methods. The second category includes deep learning methods, in particular convolutional neural networks. The recent success of 2D-CNN (Krizhevsky et al. 2012a) in image classification has motivated researchers to employ these models for action recognition in video. Several models, e.g., 3D-CNN (Ji et al. 2010), two stream convolutional networks (Simonyan and Zisserman 2014), I3D (Carreira and Zisserman 2017) and Slowfast (Feichtenhofer et al. 2019) have achieved state-of-the-art performance on benchmark datasets. These models are computationally expensive, we found no studies evaluating these methods for strength and conditioning exercise classification, under specific application constraints. It is also not clear how well these methods work on real use cases.

3.3 Human pose estimation

Pose estimation refers to recognizing the postures of humans by detecting the body parts from images. It is considered one of the hardest problems in computer vision due to challenges such as occlusion, complex motion dynamics, interactions and background (Cao et al. 2019; Papandreou et al. 2017; Huang et al. 2017). Traditional

approaches (Andriluka et al. 2009; Gkioxari et al. 2013; Sapp and Taskar 2013; Dantone et al. 2013) were based on extracting handcrafted features. Current methods based on deep learning architectures (Cao et al. 2019; Papandreou et al. 2017; He et al. 2017; Newell et al. 2017) have achieved remarkable results on this task.

Recent approaches include methods such as OpenPose (Newell et al. 2017; Insafutdinov et al. 2016; Cao et al. 2019), which work by first finding the body joints and associating them using affinity fields, and DeepCut (Pishchulin et al. 2015), which uses a partitioning and labeling formulation of a set of body-part hypotheses generated with CNN-based part detectors.

OpenPose can detect and track multiple body points in real-time and with high accuracy. The most recent version of OpenPose (Cao et al. 2019) can detect 25 body parts in an input image in under 1 second with average accuracy ranging from 75.6 to 79% on recent 2D pose estimation benchmarks.

3.4 Multivariate time series classification

Time series classification is a form of supervised classification where the data is ordered. For multivariate time series classification (MTSC), each sample has multiple dimensions and a class label.

We can group existing methods for MTSC into five broad categories (Ruiz et al. 2021; Dhariyal et al. 2020): distance-based, feature-based, ensemble based, linear models and deep learning. These methods have mostly been evaluated on the UEA MTSC (Bagnall et al. 2018) archive, which contains 30 multivariate datasets. Among the methods evaluated, linear classifiers and deep learning methods have achieved high accuracy, with low running time and excellent scalability, hence we focus on this subset here.

Linear classifiers ROCKET (Dempster et al. 2019b) (RandOm Convolutional Kernel Transform) is the current state-of-the-art for both univariate and multivariate TSC in terms of accuracy and scalability. It uses a large number of random convolutional kernels in conjunction with a linear classifier. MINIROCKET (Dempster et al. 2021) is a recent extension of ROCKET. It is deterministic, faster and more efficient than ROCKET. Unlike ROCKET, MINIROCKET implicitly normalizes the time series, thus it is scale invariant, which for our application proves to be a weakness; for MP the magnitude of the signal plays an important role in the classification task and thus the signal should not be normalized.

Deep learning classifiers Recent success on image classification (Simonyan and Zisserman 2015; Krizhevsky et al. 2012b) has motivated researchers to use deep learning methods to classify time series data. Fawaz et al. (2019) presented a comprehensive study of 9 deep learning models to classify univariate and multivariate time series. Fully Convolutional Networks (FCN) and ResNet have shown state-of-the-art performance without suffering from high time and memory complexity.

4 Data collection

Crossfit workout dataset The data used for evaluating our approach consists of video recordings of the execution of the Military Press exercise (MP). During this exercise the barbell is lifted to shoulder height and then smoothly lifted overhead by extending the elbows. The amount of weights lifted and time taken for each repetition may vary from participant to participant. MP is an important exercise in strength and conditioning, injury risk screening, and rehabilitation (Whelan et al. 2016).

Participants were asked to complete fixed repetitions of normal and aberrant forms for this exercise.

Figure 2 shows some examples for the execution of the MP exercise.

Participants 53 healthy volunteers (31 males and 22 females, age: 26 ± 5 years, height: 1.73 ± 0.09 m, body mass: 72 ± 15 kg) were recruited for the study. Participants did not have a current or recent musculoskeletal injury that would impair performance of multi-joint upper limb exercises. The Human Research Ethics Committee at University College Dublin approved the study protocol and written informed consent was obtained from all participants before the study start.

Experiment protocol The testing protocol was explained to participants upon their arrival at the laboratory. Participants completed 10 repetitions of the normal form and 10 repetitions with induced deviations. In order to ensure standardization, the technique was considered acceptable if it was completed as defined by the National Strength and Conditioning Association (NSCA) guidelines. The induced forms were chosen based on common deviations listed in the NSCA guidelines (Baechle and Earle 2008) and through discussion with sports physiotherapists and strength and conditioning coaches. Participants were allowed to familiarize themselves by completing practice repetitions.

All the performances were observed and labelled by an expert. If the performance had degraded due to fatigue then this would have resulted in the data being excluded at source. Each repetition was observed and if any repetition was not consistent with

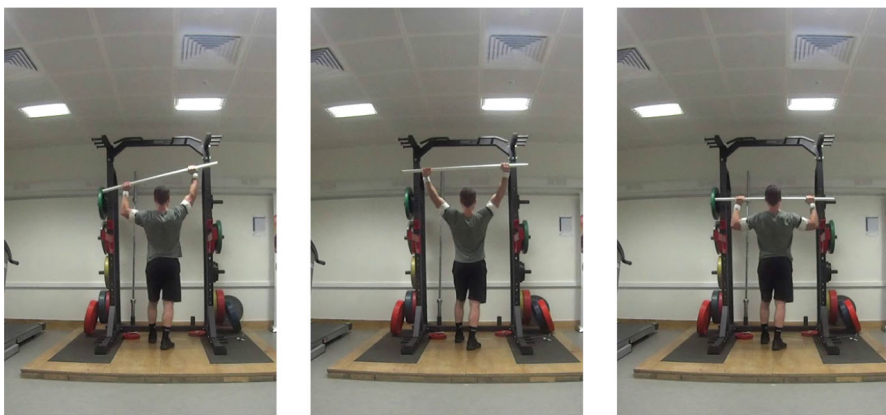


Fig. 2 Single frames depicting the induced MP deviations for class A, Arch and R (left to right)

the label (based on domain expert observation) it would have been excluded from the data.

Two cameras (Sony Action Camera, Sony, Tokyo, Japan) were set up in front and to the side of the participants to allow for recording in the frontal and lateral planes simultaneously. The data is recorded at a rate of 30 frames per second with 720p resolution. Each of these individual video clips were then labeled according to participant number, exercise completed and if they were completed in an acceptable or aberrant manner. Each participant completed the set at their desired tempo.

Exercise technique and deviations The induced forms were further sub-categorized depending upon the exercise. The deviated execution forms undermine the performance of the participants leading to a higher chance of injury. Completing the exercises with this aberrant technique means strength gains are not made as efficiently and can increase the likelihood of injury. Below we describe the four classes of normal and deviated execution forms for the MP exercise.

Exercise Classes. **Normal (N):** This class refers to the correct execution of the exercise. The participant starts by lifting the bar from near shoulder to all the way above the head until the arms are fully stretched and then bringing it back to shoulder level with no arch in the back. The bar must be stable and parallel to the ground and the back should be straight. **Asymmetrical (A):** This form refers to the execution when the bar is lopsided and asymmetrical. **Reduced Range (R):** This form refers to the execution when the bar is not brought down completely to the shoulder level. **Arch (Arch):** This type of execution indicates that the participant arches their back.

5 Methods

In this section we present the BodyMTS pipeline and provide details about individual components and data pre-processing steps. We give a description of OpenPose and why it is considered above other pose estimation libraries. We also briefly explain the process of obtaining multivariate time series data from videos using OpenPose, segmenting the long time series to obtain individual time series for each repetition and the methods chosen for the classification.

5.1 Methodology

BodyMTS is a novel end-to-end approach to classify video-based S&C exercises. It consists of two main steps. The first step applies human pose estimation to extract multivariate time series data from video. The time series data is obtained by applying pose estimation, which tracks the location coordinates of multiple body parts over the video frames. The second step applies multivariate time series classification methods.

Body pose estimation We select OpenPose (Cao et al. 2019) over other frameworks such as R-CNN (Girshick et al. 2013) or Alpha-Pose (Fang et al. 2017), for the following reasons: (1) it is robust against possible occlusions including during human-object interaction; (2) it is a full-fledged system and does not require manual steps such

Table 2 The 25 body parts tracked by OpenPose (v1.7) in a video frame (Cao et al. 2019)

0 Nose	8 Mid hip	16 Left eye
1 Neck	9 Right hip	17 Right ear
2 Right shoulder	10 Right knee	18 Left ear
3 Right elbow	11 Right ankle	19 Left big toe
4 Right wrist	12 Left hip	20 Left small toe
5 Left shoulder	13 Left knee	21 Left heel
6 Left elbow	14 Left ankle	22 Right big toe
7 Left wrist	15 Right eye	23 Right small toe
		24 Right Heel

as generating all frames for videos, setting up a display to visualize the results and saving the results in a desired format; (3) it can run on different platforms, including Ubuntu, Windows, Mac OSX, and embedded systems; (4) the inference time of OpenPose outperforms all state-of-the-art methods, while preserving high-quality results. OpenPose also provides a confidence score for each body part and each frame. The confidence score ranges from 0.0 to 1.0, where a higher confidence value indicates a higher probability of detecting a particular body part in that location. This score can be utilized as a proxy to assess the accuracy of OpenPose.

Table 2 shows the list of body parts tracked by OpenPose. It outputs the X , Y coordinates and detection confidence, for the detected body parts in a given input frame.

The coordinates are obtained with origin as a reference at the top left side of the image. The video is fed to OpenPose to obtain a sequence of X and Y location coordinates for each body part and each video frame.

Each frame is considered a single time point in the output time series data. The original videos require 813 MB of storage and after cropping and removing the audio this reduces to 213 MB. After applying OpenPose and extracting the body parts time series, the data size reduces to 30 MB, roughly a reduction of 7 times. Figure 3 shows the use of pose estimation to track the coordinates of body parts over all the video frames. The plot at the end shows the raw Y -coordinates for 8 upper body parts for a single video from class N . The time series obtained for lower body parts such as ankles, hips, etc., do not show much variability throughout the whole clip as these body parts mostly remain static throughout the execution of the Military Press.

Multivariate time series data Each video records 10 repetitions for each exercise execution, resulting in a time series capturing the body points movement for 10 repetitions. Since each repetition is the record for a single exercise execution, segmentation of the long time series is required to obtain the sequence for a single repetition. Each repetition forms a single time series sample for training and evaluating a classifier. We use peak detection methods to segment the pose estimation time series data. We find the location of local maxima in the signal using the `scipy` package.³ The body parts that are considered for finding the peaks are elbows or wrists, as these are the only body parts showing regularity in the patterns as shown in Fig. 3. We only keep the upper

³ <https://github.com/scipy/scipy>.

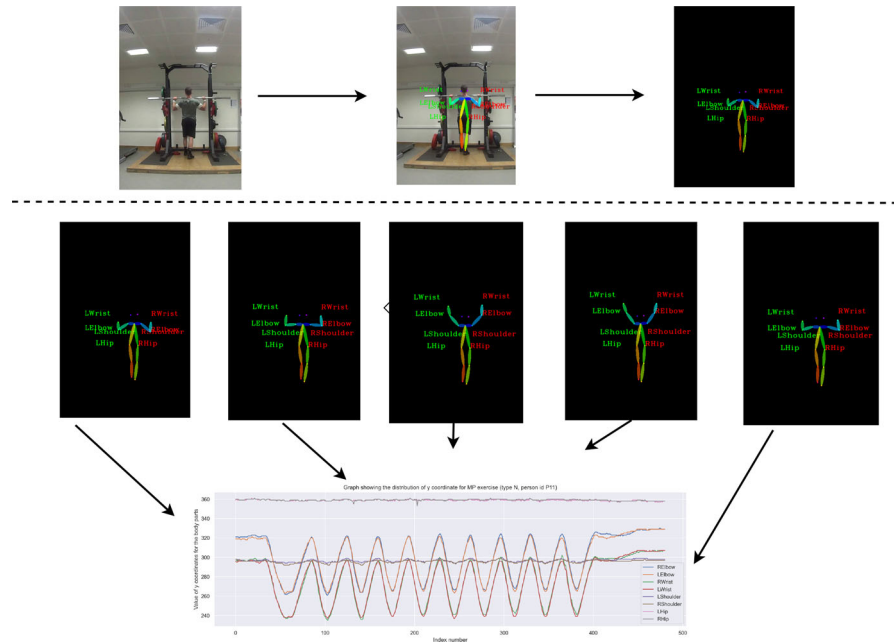


Fig. 3 Extraction of time series data from video using OpenPose. Each frame in the video is considered as a single time point in the resulting time series. Each tracked body part results in a single time series that captures the movement of that body part. The whole motion is captured as a multivariate time series with 50 channels, two (X, Y) channels for each body part tracked (only 8 body parts with Y coordinate shown above). A class label is associated with each such multivariate time series

body parts time series as suggested by the domain experts who carried out the data collection.

The data for some body parts (nose, eyes) are ignored as OpenPose fails to track these since the participant is not facing the camera. We also present results with different subsets of dimensions to understand the impact of different body parts on accuracy. We investigate using all the channels or using automated channel selection methods (Sect. 7).

The time series obtained after this step have variable length since the time taken to complete each repetition differs from participant to participant. Since the current implementation of ROCKET and deep learning methods cannot handle variable-length time series, all time series have been re-sampled to the length of the longest time series (with a length of 161). We use a 1D interpolation function with a cubic spline fit to interpolate each time series to same length (Virtanen et al. 2020).

The final data contains time series corresponding to 8 body parts (elbows, shoulders, wrists and hips) with 16 channels (X and Y coordinates). Lastly, as observed in our prior work (Singh et al. 2020) the time series data is not normalized as normalization leads to a substantial drop in accuracy. Through this application we learned that most state-of-the-art time series classifiers only work with fixed-length time series and also have an implicit step of normalizing the time series. While these algorithmic constraints

Table 3 Total number of samples per class in the train and test datasets for one 70:30 split

Class	Training	Test	Total
N	370	150	520
A	360	150	510
R	361	151	512
Arch	361	150	511
Total	1452	601	2053

seem harmless for clean TSC benchmarks, they prove problematic for real use cases. Re-sampling the length changes the meaning of the time series, and similarly, default or implicit normalization done within the algorithm, changes the meaning of the data and affects the accuracy of the classifier.

Among the classifiers evaluated, only ROCKET exposes the option of data normalization to the user, and this makes a 10 percentage points difference in accuracy for our application.

Train/test split We perform repeated 70:30 splits on the full data set to obtain training and test data. Each split is done based on the unique participant IDs to avoid leaking information into the test data. By splitting on the ID level we make sure that all the samples from a particular participant go into either the training or test data. The data is overall balanced across all the classes. Table 3 shows the number of samples across the four classes, for a single train/test split. There are roughly 1400 and 600 samples in training and test data respectively.

6 Data mining challenges and solutions in the context of BodyMTS

In this section we present the data mining challenges posed in the context of this application's requirements. We discuss the challenges and solutions for each stage of BodyMTS as shown in the Fig. 1.

- *Data size* Video data is large and requires extensive memory and storage resources as well as high end computation machines.

Solution: We investigate approaches to reduce the video size, such as frame cropping and centering, increasing the video compression ratio (using CRF) and reducing the video to time series. We find that there are settings that give a significant data reduction (70%), and still preserve classification accuracy. Experiments investigating this challenge are presented in Sect. 7.2.

- *Noisy data* The steps of video data capture and reduction, pose estimation or time series pre-processing can reduce the quality of the data by introducing noise. BodyMTS as well as deep learning methods that work directly with video are affected by the level of data noise (e.g., blurred, poor quality videos). The accuracy of OpenPose is also directly impacted by video quality.

Solution: We examine the impact caused by data reduction on accuracy. We evaluate different settings for training and predicting on high quality data, as well as training on high quality, and predicting on lower quality data (more noise) to simulate realistic use cases. We also evaluate different settings for pose estimation

and time series processing. We find that there are data and model settings that are robust to noise, preserve classification accuracy and have near real-time prediction. Experiments investigating this challenge are presented in Sect. 7.2.

High dimensionality of multivariate time series and scalability of existing classifiers Pose estimation libraries such as OpenPose track multiple key-points on a human body (25 for OpenPose), which may lead to large scale multi-dimensional time series data.

Solution We overcome this challenge by consulting domain experts as well as evaluating recent approaches to automatically select useful channels. Out of 25 body parts, we show that involving only 8 upper body parts achieves an accuracy of 87% on average for Military Press. We also explored techniques such as skipping frames in the input video during pose estimation. We examine recent scalable multivariate time series classifiers and find that ROCKET behaves the best among state-of-the-art methods, with regard to both accuracy and training/prediction time. Experiments investigating this challenge are presented in Sect. 7.3.

- *Segmentation of time series* The input video is a sequence of 10 repetitions of one exercise, which simulates actual use cases. The classification is done on each individual repetition, thus a segmentation step is required to break down the data into single reps. The current method utilizes peak detection on pose estimation time series to obtain data for each repetition. This approach is prone to noisy fluctuations (due to pose estimation errors) and may lead to incomplete repetitions. Additionally, it is not known in advance which body parts to utilize for segmentation.

Solution: We investigate different channel selection methods to capture the subset of body parts that is more relevant for the target exercise. We also analyse simple segmentation techniques directly on video, versus segmentation based on pose estimation time series data. We find that directly splitting the video into equal parts works reasonably well, although it is less accurate than using peak detection on pose estimation time series. Experiments investigating this challenge are presented in Sect. 7.1.3.

- *Data privacy and security* Capturing video data comes with implicit challenges such as privacy and data security. It becomes absolutely critical to design approaches that maintain the individual anonymity.

Solution: We turn videos into pose time series, which removes any visual clues about the identity of users. Pose estimation could in principle be used for height estimation, but this reveals little about the identity of users.

7 Experiments

This section presents an empirical evaluation of our method and is organized around the data mining challenges discussed in the previous section. In Sect. 7.1 we compare the performance of different deep learning methods for video classification versus BodyMTS with regard to accuracy and computational efficiency. We also evaluate the impact of two segmentation techniques on the performance of BodyMTS and the best

deep learning method. We further compare the impact of video quality on the best deep learning method versus BodyMTS. Section 7.2 addresses challenges raised by noisy data and data size. We analyze the robustness of BodyMTS against different sources of noise which can be broadly grouped into 3 categories: (1) video data capture; (2) OpenPose parameters; (3) time series data pre-processing.

Section 7.3 addresses high dimensionality and scalability for time series classification. We present the accuracy and compare the total execution time for different classifiers. We further evaluate the impact of utilizing different subsets of body parts on BodyMTS accuracy. The best results are highlighted in bold in the tables.

We have not included any experiments to address the issue of data privacy and security as BodyMTS works directly on the multivariate time series data and hence reasonably safeguards the identity of the user.

7.1 BodyMTS versus direct human action recognition classifiers

In this section, we compare BodyMTS with state-of-the-art methods for human activity recognition from videos. These methods employ deep learning architectures and have shown good performance on several benchmark datasets such as UCF101, Kinetics-400 and Kinetics-600. We selected a few methods based on their performance, execution time and resources required. The following section provides a brief overview of the selected methods.

- *C2D* (Fan et al. 2020) stands for 2D convolutional based model. All convolutions are performed on each of the frame. ResNet-50 and ResNet-101 can be used as the backbone architectures for C2D.
- *I3D* (Carreira and Zisserman 2017) stands for inflated 3D Convolutional network. They work by inflating the kernels of C2D models in order to capture the temporal information. These models are computationally expensive due to an increased number of computations.
- *SlowFast* (Feichtenhofer et al. 2019) architecture uses two pathways to perform activity recognition from videos. The slow pathway captures the spatial information at low frame rates and the fast pathway captures the temporal information at high frame rates. The fast pathway requires less computation because it uses a backbone network with reduced channel sizes, which is normally 8 times smaller than slow path backbones. The information from both the pathways is fused by lateral connection. The backbone architecture of SlowFast can be a 3D ResNet-50, 3D ResNet-101, a Non-local Network or a combination of these.
- *Non-local Network* (Wang et al. 2017) is used to capture the long range dependencies by enhancing the large receptive fields. They can be integrated as a generic building block with most deep learning architectures. In the experiments by (Feichtenhofer et al. 2019) they have been combined with existing standard backbone models such as ResNet-50 or ResNet-101.
- *X3D* (Feichtenhofer 2020) progressively expands a 2D CNN along multiple axis in space, time, width and depth. It uses the progressive forward expansion followed by backward contraction. The axis is selected based on the performance of the model.

Table 4 Selected deep learning activity recognition models and their configurations

Model name	Model config	No. of frames	Sampling rate	Size
C2D	C2D_8×8_R50	8	8	224 ²
I3D	I3D_8×8_R50	8	8	256 ²
SlowFast	SLOWFAST_4×16_R50	32	2	256 ²
SlowFast + NL	SLOWFAST_NLN_4×16_R50	32	2	256 ²
X3D-M	X3D_M	16	5	256 ²

7.1.1 Deep model architectures

We use the SlowFast library (Fan et al. 2020) to evaluate the above mentioned models. Table 4 shows the number of frames, sampling rate and the frame cropping size for these architectures. The column “Model Config” refers to the name of the config file in the SlowFast repo.⁴ All models are initialized with the weights pre-trained on Kinetics-400. In column “Model config”, *R50* indicates the ResNet-50 has been used as the backbone architecture; the numeric values in the format $X \times Y$ where X indicates the number of frames and Y indicates the sampling rate, e.g. C2D_8×8_R50 is a C2D model which utilizes a total of 8 frames with sampling rate of 8 from a video clip. All the experiments are executed on an Ubuntu machine with a single GPU (NVIDIA TITAN XP 12 GB, Ubuntu 18.04.5 LTS, AMD Ryzen 7 1700X Eight-Core Processor). We evaluate different values of batch size to utilize the maximum amount of GPU. After multiple iterations we found a batch size of 5 suitable for all the architectures to avoid getting out of memory errors.

We use 10% of training data as validation set. The validation set is chosen such that participant ids of training and validation are not overlapping. Table 4 shows the selected models: SlowFast, C2D, X3D, and I3D. In the case of testing, 10 frames are uniformly sampled from each clip consisting of a single repetition along the temporal axis. Each frame underwent top/left, middle and bottom/right cropping thus giving a total of 30 frames for each video. The final softmax score is averaged over the 30 frames to give the final prediction. We report the training and testing time for each of these architectures. Please refer to (Feichtenhofer et al. 2019) for more details on the data pre-processing and configurations. Note that directly running these models with the default parameters given in the SlowFast repository does not lead to good results. A considerable amount of engineering effort was spent in tuning the hyperparameters: learning rate, batch size, epochs, warm-up epochs and weight decay. All other hyperparameters remain unchanged. All the models are trained with Stochastic Gradient Descent (SGD) with momentum 0.9 for 10 epochs.

7.1.2 Results of BodyMTS versus deep learning models

Table 5 reports the average accuracy and average running time of the deep learning models evaluated. Note that this accuracy was obtained with significant model engi-

⁴ <https://github.com/facebookresearch/SlowFast>.

Table 5 Average accuracy, total testing time, time per testing clip and total training time for different architectures over 3 train/test splits

Model	Accuracy	Total test time (mins)/test time per clip of 10 reps (mins)	Training time (mins)
C2D	0.67 (± 0.021)	23/0.38	19
I3D	0.79 (± 0.036)	32/0.53	23
SlowFast	0.83 (± 0.021)	29/0.48	52
SlowFast + NL	0.83 (± 0.020)	27/0.45	59
X3D-M	0.78 (± 0.012)	41/0.68	126
BodyMTS (frame-step=1)	0.87 (± 0.026)	22/0.36	52
BodyMTS (frame-step=3)	0.85 (± 0.029)	12/0.20	26

The average duration of all clips in training and testing is 65 mins and 30 mins respectively. All deep learning models are pre-trained on Kinetics-400. For accuracy without pre-training, see Table 6

neering effort and with using pre-trained weights from the Kinetics-400 benchmark for all the deep models. As seen from Table 5, C2D performed the worst whereas SlowFast and SlowFast + NL achieve the highest accuracy, followed by I3D and X3D-M. We select the best architecture model which is SlowFast in this case and compare its performance with BodyMTS. We use the short name SlowFast for this model in the subsequent sections. In the test data there are a total of 60 clips each with 10 repetitions on average. These clips were fed one by one to the pre-trained model. The testing time for each model is the summation of total time taken for data pre-processing (which includes segmentation), model loading and classification. Table 5 shows the total testing time taken for 60 clips for each model. We also reported the average testing time over single clips of 10 repetitions. The training time shown is the summation of time taken for data pre-processing and model training for all the clips in the training data.

Accuracy Table 6 reports the average accuracy over three splits for SlowFast (0.83) and BodyMTS (0.87). BodyMTS achieves higher accuracy with minimal model engineering effort. We note that the default architecture of SlowFast is meant for classifying large scale video datasets such as UCF-101 or Kinetics-600 and so there is a higher chance of overfitting. This is substantiated by observing that there is a very significant drop in the accuracy of SlowFast when removing the pre-trained weights as shown in Table 6, with the accuracy dropping by 60 percentage points, from 0.83 to 0.25. It is only after using the pre-trained weights on Kinetics-400 and other model engineering steps that SlowFast reaches an average accuracy of 0.83.

Execution time We report the total training and testing time for both the models in Tables 5 and 6. The total duration of all the videos (both training and test) is 95 min. We observe that the combined train/test time of BodyMTS (OpenPose + data pre-processing + training/testing) is around 74 min whereas the combined train/test time of SlowFast is around 86 min. Prediction time is 22 min for BodyMTS and 27 min for SlowFast. This shows that BodyMTS is faster than SlowFast for both training and prediction. Additionally, using a frame step of 3, the combined train/test time of

Table 6 Comparison of SlowFast and BodyMTS approaches in terms of time taken and resources required

Step	SlowFast	BodyMTS (frame-step = 3)
Data size	213 MB (videos)	28 MB (time series)
Training time	59 mins	26 mins
Testing time	27 mins	12 mins
Average accuracy over three splits	0.83 (with pre-training on Kinetics-400, epochs = 10) 0.25 (no pre-training, epochs = 50)	0.85
Infrastructure	GPU	GPU or CPU

The total duration of all clips in training and testing is 65 mins and 30 mins respectively

BodyMTS goes down to 38 min which is significantly faster than the time taken for SlowFast. Excluding the execution time of OpenPose, BodyMTS only takes a total of 2 min including both the training and testing time for the classifier. Additionally, for a single clip consisting of 10 repetitions BodyMTS takes a total time of 12 s, whereas SlowFast takes 29 s on average. Hence BodyMTS is overall faster than SlowFast and can deliver near real-time predictions.

Cost All deep learning methods require the use of GPU. In our case, a single GPU was utilized for the deep learning models to compare their performance. Increasing the number of GPUs will lead to a reduced execution time but at the cost of expensive infrastructure. However, running deep learning models without any GPU will lead to significant increase in training/testing time. The CPU version of OpenPose takes roughly about 15sec/frame (Cao et al. 2019), however, the recent lightweight implementations of OpenPose makes it possible to reach real-time inference on CPU with negligible accuracy drop (Osokin 2018). We recently found that libraries such as OpenVINO⁵ makes it possible to execute OpenPose on the CPU machines and TensorFlow Lite⁶ supports running pose estimation models directly on the mobile phone in real-time.

This is interesting for future work, but a different workflow than the one we use here. Therefore, it is possible to reduce the computation footprint of BodyMTS more than the corresponding deep learning models for video-based exercise classification.

Storage space Table 6 shows the initial data size for SlowFast and BodyMTS. Because of the large size of videos there is high cost involved to store the videos. However, videos do not need to be stored for BodyMTS, once the time series are extracted. Since the data is just a sequence of numbers (time series), there can be large savings in storing this data.

Even when the data is sent to the cloud server, there can be large savings in terms of bandwidth required to transfer the data.

⁵ https://docs.openvino.ai/2019_R1/_human_pose_estimation_0001_description_human_pose_estimation_0001.html.

⁶ https://www.tensorflow.org/lite/examples/pose_estimation/overview#performance_benchmarks.

Table 7 Average accuracy obtained by SlowFast and BodyMTS for two different segmentation approaches for three train/test splits

Segmentation approach	SlowFast accuracy	BodyMTS accuracy
With pose estimation	0.83	0.87
Without pose estimation	0.77	0.80

Practical aspects From Tables 5 and 6, we see that BodyMTS has higher accuracy than the deep learning methods. Deep models may suffer from high training and test time depending upon the data size as well as require high engineering effort to tune hyper parameters such as the number of epochs, learning rate, etc. Increasing the number of GPUs may lead to decrease in execution time for the deep learning models, however, this may not be always possible due to cost. In contrast, due to the lightweight nature of the time series classifier, BodyMTS does not require GPU resources (using lightweight OpenPose) and can be trained/tested within a fraction of the runtime of SlowFast, on a single CPU machine (Dempster et al. 2019b; Osokin 2018).

7.1.3 Impact of segmentation

In this section we analyze the impact of two segmentation techniques on the accuracy of BodyMTS and SlowFast. As stated earlier in the data pre-processing step, segmentation of the video data is required to obtain the individual sample for each repetition for train/test data. We consider two scenarios here:

- When the start and the end time of each repetition is known in advance. This information can be utilized to easily segment the video data into individual repetitions. To get this information we use pose estimation and peak detection techniques.
- When the number of repetitions are known in advance. Dividing the total duration of the video clip by the total repetitions can approximately segment the individual repetition assuming that the participant takes consistent time to complete each repetition. This approach does not require pose estimation information.

Table 7 shows the average accuracy of SlowFast and BodyMTS using two different segmentation approaches over three train/test splits. BodyMTS achieves 4 percentage points higher accuracy than SlowFast when the data from pose estimation is utilized for segmentation. When the segmentation is done by equally dividing the total duration by the total number of repetitions, BodyMTS still achieves 3 percentage points higher accuracy than SlowFast. Therefore, from the above results accuracy obtained is higher when the repetitions are correctly segmented using pose estimation than the scenario where the segmentation is performed by equally dividing the total duration with the total number of repetitions. This suggests that repetitions may not be fully captured as this is just an approximate way to obtain each repetition. The idle time at the start or the end of the video clip as well as the variation in the duration of each repetition can affect the segmentation.

Table 8 Average accuracy obtained by SlowFast and BodyMTS for varying video quality (CRF from 16 to 34) over 3 train/test splits

CRF	SlowFast accuracy	BodyMTS accuracy
16	0.84	0.87
23 (default)	0.83	0.87
28	0.82	0.85
34	0.81	0.81

7.1.4 Impact of video quality noise

We now analyze the impact of video quality noise on the deep learning models, as well as BodyMTS. We do this by changing the CRF video property as discussed in detail in Sect. 7.2. Higher value of CRF leads to a drop in the quality of the video and vice versa. We compare the accuracy at different values of CRF: the default is set at 23 and we test a higher quality of video at CRF 16 and degrading the quality of the video all the way to CRF 34.

Table 8 shows the average accuracy obtained with SlowFast and BodyMTS at different CRF over three train/test splits. We observe that at the default value of CRF 23, BodyMTS achieves higher accuracy than SlowFast by 4 percentage points. Reducing the video quality by increasing the CRF affects both methods, with the accuracy decreasing but still above 80% which is desirable as described in Sect. 2.

For future work, it would be interesting to study whether video quality metrics such as VMAF Aaron et al. (2015) could also be used to identify an application specific threshold beyond which the video quality is too poor for inclusion in this task. In the Appendix, we investigate a few video quality metrics and the corresponding BodyMTS accuracy.

Takeaway The previous experiments suggest that BodyMTS is more accurate, significantly faster and more cost-efficient when compared to the best deep learning method, SlowFast.

7.2 Robustness analysis: impact of noise on BodyMTS

In this section, we analyze the robustness of BodyMTS against different sources of noise that may occur in this application. These sources of noise can be broadly classified into 3 categories: (1) video data capture; (2) OpenPose parameters; (3) time series data pre-processing.

While studying the impact of noise we address the following questions:

1. How does the noise from different sources such as video capture quality, OpenPose estimation and data pre-processing affect the classifier accuracy?
2. Is it possible to reduce the quality of videos but still keep the same accuracy? Are there possible benefits in terms of saving storage space by reducing the data size?

We address these questions by generating different types of noisy videos with varying levels of noise. We degrade or enhance the quality of videos by changing the resolution and bit rate. Also, we analyze the impact of OpenPose parameters that

are most responsible for affecting the quality of estimation. In addition, we explore parameters of OpenPose that can be tuned to reduce the overall execution time of BodyMTS. We further analyze how much the quality of videos can be reduced by changing the compression level without sacrificing accuracy.

We consider two scenarios: 1. Adding noise to both training and testing data. 2. Adding noise to testing data and keeping the training data intact.

7.2.1 Data-capture noise

In this section, we study the impact of noise coming from data capture. This can be further categorized into two types: video quality and the recording conditions. We describe each of them in detail below.

Video quality The motivation behind studying the video quality is that videos captured in the wild can range from very poor quality to high definition quality. For modern smartphones the camera quality is much more efficient than it used to be 10 years ago. Nevertheless, recorded videos can still have a low quality because of the compression required to send the data to the cloud service where further processing takes place. We generate varying quality of video data with different levels of noise by tweaking the video properties such as the bit-rate and resolution. We use FFmpeg Tomar (2006) to obtain noisy videos by modifying the above properties. This is an open-source, widely used tool used for manipulating and modifying videos.

Additionally, OpenPose estimation confidence is directly proportional to the quality of the video. Lower quality videos lead to low confidence of the body-parts location estimation, which ultimately may affect the classifier accuracy.

- *Resolution* refers to the number of pixels in each image. The higher the resolution, the more pixels and hence the better video quality. In our experiments, we downscale the original resolution of the videos to measure the impact on classifier accuracy. Note that the original videos of the Military Press were recorded at HD quality with 720p resolution.
- *Bit-rate* refers to the amount of information processed per second to represent a video. Higher bit-rate means more information, which means better quality, but also a higher video file size. We use the constant rate factor (CRF) Tomar (2006), which is a rate control mode, to change the number of bits per frame. We examine the impact of bit rate by changing the CRF property of videos. Higher CRF leads to lower quality videos.

In addition to the above properties, we also alter the color space and frame rate to analyze their impact on the BodyMTS accuracy.

Recording conditions Several factors during the video recording can affect the way a video is recorded which ultimately can affect the confidence of OpenPose. We categorize common recording conditions into different types and provide a brief description below.

- *Camera settings* This may include factors such as orientation, viewing angle, zoom of the camera, distance between the participant and the camera or whether the participant is centralized or not, etc.
- *Background* Depending upon the conditions such as whether the participant is executing the exercise in an indoor or outdoor settings, this may influence the quality of video. For instance, in a gym setting factors like multiple people, background noise (pictures having humans), clothing (background color same as clothing) and lighting may affect the final output of videos.

Apart from the above, there may be other unaccounted factors which can influence the recording conditions. In our data, the distance to the camera and the background can change, and these variations did not affect the accuracy of BodyMTS. Nevertheless, we note that BodyMTS expects certain conditions to be met (e.g., regarding camera being stable) as listed out in Sect. 2 before deploying.

7.2.2 OpenPose parameters

BodyMTS utilizes OpenPose to obtain the coordinates information for major keypoints in a human body. OpenPose is a full-fledged, 2D pose estimation system that includes many parameters that affect accuracy, optimization, display and output format. Here the objective is to evaluate and tune the parameters which may influence the accuracy and efficiency of OpenPose which ultimately affects the running time and accuracy of BodyMTS. A short description of these parameters is provided below:

- *Frame-step* an integer value indicating the number of frames to skip during the estimation.
- *Net-resolution* increasing this may increase the accuracy while also increasing the execution time.
- *Scale-number* this parameter indicates the number of scales to average.

The parameters *net-resolution*, *scale-number* are directly responsible for the accuracy of OpenPose.

OpenPose fails to detect the body parts in case the person is not facing the camera or the body part is cropped accidentally or the camera fails to capture the whole body. In the Military Press execution, the participant is not facing the camera and so OpenPose fails to detect the coordinates for eyes and nose. Nonetheless, these body parts are not involved in the physical movement and so have no impact on the accuracy.

All the experiments are performed using the sktime (Löning et al. 2019) version of ROCKET on an Ubuntu 18.04 system (16 GB RAM, Intel i7-4790 CPU @ 3.60 GHz).

7.2.3 Results for the impact of video quality on BodyMTS

In this section we analyze the impact of changing the video quality on the BodyMTS accuracy. We start by changing the bit-rate and resolution both of which are responsible for determining the quality of video. We further present the total size of the video data

Table 9 Average accuracy of BodyMTS on test data over three train/test splits for different video frame resolution

Resolution	Total size of videos (MB)	BodyMTS accuracy
Default (420 × 460)	213	0.87
One-half (210 × 230)	48	0.82
One-third(140 × 154)	27	0.75

Table 10 Average accuracy of BodyMTS on test data over three train/test splits for different values of CRF

CRF	Total size of videos (MB)	BodyMTS accuracy
Default (23)	213	0.87
16	398	0.87
22	208	0.87
28	76	0.85
34	34	0.81

At CRF 28 we save 70% of data storage and maintain similar accuracy

obtained after altering each property. We note that in the following experiments we consider the case when both the train and the test data have been impacted by the changes in video quality.

Reducing the resolution We reduce the original resolution in steps of one-half, one-third of the original resolution and evaluate its impact on the classifier accuracy. Table 9 shows the impact of various resolutions on the classifier accuracy and data size.

Reducing the bit-rate We alter the CRF in order to modify the bit rate. CRF ranges from 0 to 53 and the default value of CRF is 23. We change the value of CRF with a step size of 6 as suggested in (Tomar 2006), starting from 16. Resolution remains the same when changing the CRF. Table 10 shows the impact of changing the CRF (or bit-rate) on the classifier accuracy.

Results and discussion We see in Table 9 that reducing the resolution has a negative impact on the classifier accuracy. The reduction in average accuracy is more than 10 percentage points when the original resolution is reduced to one-third. This confirms that degrading the quality of videos by reducing the resolution leads to a significant drop in accuracy. We are interested in finding good trade-offs between saving storage space and maintaining or only slightly sacrificing the accuracy.

Next we alter the CRF property in order to change the bit-rate. Increasing the CRF leads to a drop in the bit-rate which leads to degraded quality of the videos and vice versa. Table 10 shows that increasing the value of CRF leads to a drop in the classifier accuracy whereas decreasing this value has no effect on the accuracy. Figure 4 shows a single frame for class Normal at CRF 23 (default) and CRF 40 where the image becomes too distorted to be usable.

The change in the video quality at CRF 22 and 23 is insignificant and hence the accuracy remains consistent. We observe that setting the value of CRF to 28 has no



Fig. 4 Single frame for class Normal at default resolution of 420×460 and CRF 23 (default) versus CRF 40 (lower quality, distorted). The figure at CRF 40 (right) has a low resolution and hence looks blurrier than the figure at CRF 23 (left)

major impact on the classifier accuracy. This suggests that it is possible to reduce the total storage space of original videos while maintaining accuracy. The total size of the original videos is 213 MB at CRF 23 but it is 76 MB at CRF 28, hence a saving in storage space of 70%. Additionally, the size of final time series is 28 MB which suggests further savings in storage space as compared to the original videos.

Takeaway Degrading the quality of videos by altering CRF to 28 makes it possible to satisfy minimum accuracy requirements (e.g. above 80%) as listed in Table 1 with 70% savings in storage space.

7.2.4 Results for impact of noise due to OpenPose parameters

In this section, we analyze the impact of changing the OpenPose parameters as discussed in the previous section. Table 11 shows the accuracy for different parameter values and the total training and testing time of BodyMTS. Training time includes time taken for running OpenPose, data pre-processing and training the model, similarly testing time includes time taken for running OpenPose, data pre-processing and testing the model. Note that the impact of changing OpenPose parameters has been evaluated on the original dataset (default video settings for resolution and CRF). There are a total of 205 clips (2053 repetitions) of Military Press with a total combined duration of 1 h 35 mins.

Results and discussion From Table 11 we observe that increasing the *frame-step* from 1 (using every frame) to 3 (every third frame), leads to a small drop in accuracy of 2 percentage points, but a significant reduction in run-time of OpenPose.

We further observe that increasing the values of the parameters *net-resolution* and *scale*, which are mainly responsible for the confidence of OpenPose, produce no

Table 11 Average accuracy on test data over three train/test splits for different OpenPose parameters

OpenPose parameters	BodyMTS accuracy	Training/testing time (mins)
default (frame-step = 1, net-resolution = -1×368 , scale-number = 1)	0.87	52/22
frame-step = 2	0.86	37/17
frame-step = 3	0.85	26/12
net-resolution = 640	0.85	174/76
net-resolution = 720	0.82	234/105
scale-number = 4	0.82	140/63

The total duration of all clips including training and testing clips is 95 mins

Table 12 Average accuracy on poor quality test data at CRF 28 over three train/test splits for different OpenPose parameters

OpenPose parameters	BodyMTS accuracy
default (frame-step = 1, net-resolution = -1×368 , scale-number = 1)	0.85
net-resolution = -1×640 , scale-number = 4 scale-gap = 0.25	0.85
net-resolution = -1×720 , scale-number = 4 scale-gap = 0.25	0.84

improvement on the accuracy, but rather leads to an increase in the overall run-time and a drop in the accuracy. These results suggest that the default values of these parameters are enough to give a reasonable accuracy using OpenPose. Table 11 also shows the total running time of OpenPose. Increasing the *frame-step* makes OpenPose faster since it is not processing all the frames and so skipping some frames leads to faster execution time. The accuracy obtained by setting frame-step to 3 is still above the minimum desired accuracy of 80% listed in Table 1, which means that for pose estimation we do not need to consider every frame, and using every third frame is enough to capture the movement of body parts relevant for this classification task.

Takeaway By using a frame-step size of 3 along with default values of the remaining parameters for OpenPose, it is possible to noticeably reduce the training and testing time without a major drop in accuracy and still satisfy the minimum accuracy requirements in Table 1.

We further tried tuning the *net-resolution* on lower quality video at CRF 28. Table 12 shows the accuracy over different parameters of OpenPose for video quality at CRF 28. We observe no major improvement over the previous accuracy of 0.85. This also suggests that video quality plays a crucial role in determining the accuracy of BodyMTS as tuning OpenPose parameters alone is not sufficient to achieve good performance.

7.2.5 Training on good quality videos and testing on poor quality videos

Previous results shown in Tables 9, 10 and 12, considered the scenario when both the train and test video data are impacted by the same degree of level of noise. In this

Table 13 Average accuracy of BodyMTS on test data for different values of CRF over three train/test splits

CRF	BodyMTS accuracy	Total size of videos (MB)
Original data (CRF = 23)	0.87	213
28	0.85	76
30	0.83	56
34	0.78	34

The training is performed on original data (CRF = 23), while the testing is performed on poor quality video by altering the CRF from 23 to 34

section, we consider another scenario when training is performed using the original high quality video whereas testing is performed using poor quality video. The poor quality videos are generated by altering the CRF value. Table 13 shows the accuracy of BodyMTS when trained on high quality videos and tested on poor quality videos.

Results and discussion We observe from Table 13 that BodyMTS accuracy drops when trained on high quality videos and tested on poor quality videos after CRF 30. From the above results it is clear that a threshold of CRF 30 can be chosen to reduce the data size while still satisfying the minimum accuracy requirements as listed in the Table 1. As previously observed in Table 10 we save 70% in storage space at CRF 28 without compromising on the accuracy.

7.2.6 Discussion on impact of video quality and OpenPose parameters on BodyMTS

In the above experiments, we studied the impact of two major sources of noise from video quality and the OpenPose parameters on the accuracy of BodyMTS. We observed that the quality of the videos has a large impact on the classifier accuracy. We also found that degrading the quality of videos by introducing a small amount of noise (at CRF 28) can lead to large savings for the storage space (from 213 to 76 MB), at a very small drop in accuracy (from 87 to 85%). This can be essential for applications deployed on low memory devices such as mobile phones, as well as where bandwidth is a constraint. From the pose estimation side we observed that the default values of OpenPose parameters are sufficient for good accuracy. The total duration of the original videos is 1 h 38 min, whereas OpenPose took 1 h 12 min using default parameters and this was further reduced to 38 min by utilizing every third frame during the estimation with a very small drop in accuracy (from 87 to 85%). Thus we can save both storage space and run-time, which is very promising given the constraints and requirements of this application. We note that there are many other types of video noise we can investigate, but we focused here on a subset of the most relevant sources for this application.

7.3 Robustness analysis: time series classification in BodyMTS

In this section, we evaluate several multivariate time series classifiers and report the average accuracy over 3 train/test splits.

Table 14 Average accuracy on test data over 3 splits for multivariate time series classifiers trained with time series extracted with OpenPose version 1.4 and version 1.7

Time series classifier	Accuracy OpenPose (v1.7)	Accuracy OpenPose (v1.4)
FCN	0.82 (± 0.012)	0.72 (± 0.043)
ResNet	0.76 (± 0.040)	0.73 (± 0.028)
ROCKET	0.87 (± 0.026)	0.81 (± 0.03)
MINIROCKET	0.81 (± 0.030)	0.75 (± 1.644)

Table 15 Average time taken for training and testing over 3 splits for the selected methods

Time series classifier	Training time (mins)	Testing time (mins)
FCN	85	28
ResNet	115	28
ROCKET	52	22
MINIROCKET	50	20

Note that training and testing time shown here is inclusive of the time taken for OpenPose and data pre-processing. The total duration of all clips in training and testing is around 65 mins and 30 mins respectively

We present results using both the previous version of OpenPose (v1.4) and the latest version (v1.7 at the time of writing). We are interested to see if the improvements in OpenPose lead to a significant improvement in the classification accuracy. For FCN and ResNet, we did not tune any hyperparameters and used the defaults as mentioned in the original papers (Fawaz et al. 2019). We observe that for ROCKET (Dempster et al. 2020) changing the number of kernels (default=10000) did not produce any significant change on the accuracy, hence we kept the defaults as recommended by the authors. Where the algorithm allows it by exposing this option to the user, we disable the time series normalisation step. We show detailed results for varying data pre-processing in the Appendix.

We also present results for different subsets of time series dimensions (i.e., body parts). We utilize the ECP method (Dhariyal et al. 2021) for automated dimension selection due to the large number of possible combinations from the 25 body parts (50 dimensions). We use both the left and right parts for a single body part unless otherwise stated. We further compare this with the number of dimensions suggested by the domain experts who carried out the data collection.

Results and discussion Table 14 shows the average accuracy and standard deviation obtained on the test data over three data splits. ROCKET achieved the highest accuracy for both versions of OpenPose, followed by the deep learning models. The standard deviation values are generally small, which means accuracy generally remains consistent over different splits. We further observe that there is a notable increase of 6 percentage points in the accuracy of ROCKET transitioning from the older version to the latest version of OpenPose (v1.7 at the time of writing this paper). We think that further improvements in OpenPose and time series classifiers can further improve the performance on this task. With an average accuracy of 87%, it confirms that Open-

Table 16 Average accuracy of ROCKET using OpenPose v1.7 on test data for different subsets of body parts over 3 train/test splits

Number of body parts	Body parts	ROCKET accuracy
25 body parts	All	0.83 (± 0.028)
8 upper body parts (domain expert)	Wrists, elbows, shoulders and hips	0.87 (± 0.026)
8 body parts (Dhariyal et al. 2021)	Wrists, elbows, big toes and small toes	0.81 (± 0.012)
6 upper body parts	Wrists, elbows and shoulders	0.87 (± 0.021)
4 upper body parts	Wrists and elbows	0.85 (± 0.012)

Pose coupled with ROCKET can surpass the minimum accuracy requirement of 80% as listed in Table 1.

Table 15 shows the training time and testing time for the selected methods. We observe that MINIROCKET takes the least amount of time for training and testing followed by ROCKET and deep learning methods. However, despite MINIROCKET being faster than ROCKET in total time taken, it is less accurate than ROCKET as shown in Table 14.

Based on the above results, we choose ROCKET as the classifier in BodyMTS as it the most accurate and scalable method among the classifiers investigated. In Table 16 we observe the results using different subsets of body parts selected from the multivariate time series. We note that careful selection of the body parts has a significant impact on the accuracy of the classifier, with the upper body parts recommended by the domain expert achieving the highest accuracy among alternative subsets considered. The automated selection method proposed by (Dhariyal et al. 2021) also selects 8 body parts, but seems to be affected by data noise in selecting lower information body parts such as toes, which do not play a role in accurately classifying the movement.

8 Lessons learned and limitations of the BodyMTS approach

BodyMTS achieves an accuracy of about 87% on the Military Press video dataset. Based on these results, it is possible to use videos as an alternative to sensor-based approaches for human exercise classification. However, further work is needed to analyze the generalizability of BodyMTS by including other strength and conditioning exercises. We analyzed the robustness of BodyMTS against common sources of noise, particularly the video quality and OpenPose parameters. We observed that video quality plays a critical role in determining the classification accuracy. We showed that video quality can be degraded to a CRF value of 28 without a significant drop in accuracy, whilst achieving savings in terms of storage space. We observed that the subset of channels (i.e., body parts) has a large impact on classifier accuracy, which necessitates further investigation. We have also seen that improvements in the body pose estimation method (from OpenPose v1.4 to v1.7) have resulted in higher accuracy, which is encouraging. Furthermore, in case of data pre-processing, segmentation plays a crucial role, as incorrect peak detection due to a noisy signal may lead to incorrect capturing of a repetition, which in turn affects the classifier accuracy. Lastly,

due to the choice of ROCKET as a classifier, BodyMTS does not involve tuning too many parameters. Hence BodyMTS as an application requires setting fewer parameters. Finally, we compared BodyMTS with state-of-the-art deep learning methods and observed that BodyMTS achieves better accuracy while requiring fewer computation resources. Additionally, a lot of engineering effort is required to tune the deep learning models.

Below we discuss some of the limitations and mitigations for this approach.

- *Types of exercises and pose estimation* BodyMTS may fail in case of exercises where the participants may have to lie down, e.g., for exercises like push-ups and sit-ups. In such cases, videos may fail to capture the full motion or the pose estimation may fail due to body parts occlusion. Further, we observed in the previous experiments (Singh et al. 2020) that the classifier struggles to classify some samples from class Normal and Arch for Military Press, which is due to the fact that the front view may not be able to fully capture the lateral motion. Techniques like extending the 2D information to 3D through triangulation techniques (Kwon et al. 2020; Nakano et al. 2020b) can be used to capture the depth information. BodyMTS is also limited by the number of body parts detected by OpenPose. Physical exercises which require tracking of certain body parts not supported by OpenPose may not be well suited for this approach. For instance, for an exercise which requires to track all vertebrae of the spine, it may not be possible with this approach as OpenPose does not currently track each vertebra of the spine. However, most types of Strength and Conditioning exercises involve majorly the upper or lower body parts movements which are tracked well with OpenPose.
- *Video data capture* Factors such as lighting, viewing angle, stability and position of the camera, camera quality, clothing of the participants, background, location, etc., can possibly influence the video recordings of the exercise. OpenPose may fail to detect some body parts if the body part is not fully captured in the video recordings. Additionally, the confidence of OpenPose is directly related to the video quality. The lower quality of the videos leads to decrease in the accuracy of OpenPose and hence low quality of final data. However, issues like video quality can be easily overcome by usage of smartphones that have a reasonable quality camera, whereas accidental cropping of body parts as well as positioning and the viewing angle can be avoided by requirements for the video recordings as mentioned in Sect. 2. OpenPose is robust against occlusions including during human-object interaction Cao et al. (2019). Moreover, lightweight versions of OpenPose (Osokin 2018) and TensorFlow Lite⁷ makes it possible to run pose estimation frameworks on resource-constrained devices such as smartphones as well as on a single CPU without sacrificing the accuracy. Lastly, recent works based on preserving the privacy and security of participants (Hinojosa et al. 2021) do not require participants to be completely visible in front of the cameras thus preserving their privacy which is in contrast to the deep learning methods that might require that participants be completely visible.

⁷ https://www.tensorflow.org/lite/examples/pose_estimation/overview#performance_benchmarks.

- *Data pre-processing* BodyMTS uses peak detection to segment the individual time series from multiple repetitions to obtain single repetitions. Due to noise, jittering and depending on the accuracy of OpenPose, segmentation may not always be correct and hence may lead to loss of some repetitions. Additionally, the duration of a single repetition cannot be generalized as it varies from participant to participant. In our experiments only a few samples were dropped due to incorrect segmentation without having a substantial impact on the accuracy.

Despite these limitations, our experiments show that BodyMTS is a very promising approach. Advancements in body pose estimation and time series classification may help in further improving the performance of BodyMTS. An interesting future direction concerns the development of time series classification algorithms that are more flexible and do not require strict pre-processing steps such as length re-sampling or normalisation. Additionally, the data produced by OpenPose has associated pose estimation confidence values, and this raises interesting research questions of how the classifier may benefit from knowledge of uncertainty in the data to improve the accuracy.

9 Recommendations for practitioners

We divide this section into three parts: video data, pose estimation and time series classification. We provide recommendations for each of these based on the experiments performed and our findings.

- *Using videos as data source for S&C exercise classification* We recommend the use of videos as an alternative source to sensors, for S&C exercise classification. The large storage and computation requirements of video can be addressed by changing the bit rate property, which reduces the video quality and size. In the case of Military Press, a CRF up to about 28 to 30, reduces the data size by 70% without affecting the accuracy.

Certain application requirements on the video quality and data capture process need to be met. For example, the requirement of a stable camera, as well as the participant being fully captured in each frame are important to ensure the quality of the follow up steps. We have detailed such requirements in Sect. 2. We recommend that the recorded videos are then pre-processed to remove any audio, background, and centralize the participants using a bounding box, followed by changing the video CRF value which results in saving storage space.

- *Pose estimation* We recommend the use of OpenPose for human pose estimation. It can process video faster than real-time, and has parameters such as *frame-rate* which allows skipping frames during processing. We have found that skipping 1-2 frames further improves speed and does not affect the accuracy of the classification step. The body parts detection and tracking is sufficiently accurate to enable the followup classification step. We found that improvements in pose estimation in v1.7 have also lead to improvements in classification.

Libraries such as OpenVINO⁸ make it possible to execute OpenPose on CPU, and TensorFlow Lite⁹ supports running pose estimation models directly on a mobile phone in real-time, so there is a lot of promise in current developments for pose estimation.

- *Data pre-processing and multivariate time series classifiers* We recommend using pose estimation information to obtain the human motion time series and then segment the multiple repetitions of the exercise. This is more accurate than other heuristics for segmenting the reps.

Before segmentation, it helps if the full signal is smoothed using a Savgol filter followed by peak detection. We recommend to not normalise the data of each repetition, as this can change the meaning of the data and also decreases the classification accuracy.

We found that length re-sampling does not impact the final classification accuracy. We recommend ROCKET for the multivariate time series classification step, as it is very fast and most accurate among the classifiers evaluated.

10 Conclusion

In this work we have analyzed the performance and robustness of BodyMTS, an approach for exercise classification using videos as time series. We presented the required features and associated technical requirements for this kind of application. We evaluated BodyMTS on a real-world dataset for the Military Press exercise and achieved an average accuracy of 87%. We further showed that the latest improvements in body pose estimation with OpenPose can improve the performance of BodyMTS. We observed that the subset of channels (i.e., body parts) has a large impact on classifier accuracy, which necessitates further investigation in future work. We compared the robustness of BodyMTS against different sources of noise particularly related to variations in video quality and OpenPose parameters. We observed that BodyMTS can achieve good performance at different levels of video quality.

We showed that by decreasing the quality of videos, a major portion of storage cost can be reduced (70%), with a very small drop in accuracy (2 percentage points). This leads to less computation as well as savings in terms of storage cost. We further noticed that changing the parameters of OpenPose has less impact on the classifier accuracy, but can lead to large savings in running time. Lastly, we compare the BodyMTS approach with deep learning-based methods for human activity recognition from videos. We considered several aspects such as performance, storage space and practicality. We observed that BodyMTS can achieve better performance than deep learning methods in terms of total time and accuracy, without the use of heavy computing resources. For future work, we plan to carry out extensive experiments to analyze how BodyMTS generalises to more strength and conditioning exercises. Additionally, we plan to compare the performance of BodyMTS with sensors-only based approaches. Furthermore,

⁸ https://docs.opencv.org/2019_R1/human_pose_estimation_0001_description_human_pose_estimation_0001.html.

⁹ https://www.tensorflow.org/lite/examples/pose_estimation/overview#performance_benchmarks.

we plan to work on the classification interpretation aspect, where the goal is to provide useful prediction explanation feedback to the end-user.

Acknowledgements This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant numbers [12/RC/2289_P2, SFI/16/RC/3835].

Appendix

Time series data pre-processing and classification

We consider the impact of time series length re-sampling and data normalization here. We use a single split of data consisting of only y coordinates of 8 body parts here for faster execution.

- *Re-sampling* Each time series has been re-sampled to the same length since most time series classifiers cannot handle variable-length data.
- *Normalization* The magnitude of the signal is important for this application, as shown in the experiment here.

Results and discussion Table 17 shows the impact of changing the re-sampling length on the BodyMTS accuracy. We see that there is almost no effect of length re-sampling on the classifier accuracy. Furthermore, reducing the length of the data also leads to reduced execution time of BodyMTS. We also experimented with the parameters of the ROCKET classifier such as number of kernels (10000) and normalization (False). While changing the number of kernels did not produce any impact on the accuracy, setting the normalization to FALSE lead to a big increase in the accuracy as shown in Table 18. We believe that this is due to the loss of magnitude information which is a key element in the classification for this type of problem. We further experimented by converting the color scale of videos to gray and observed no change in the accuracy of BodyMTS.

Table 17 Accuracy on test data over a single train/test split for different values of time series length

Re-sampling length	BodyMTS accuracy
50	0.81
100	0.83
161	0.83
225	0.83
300	0.83
500	0.83

Quantifying video quality noise using video quality metrics

We further quantify the impact of noise on the classifier accuracy by using video quality metrics. We use three scores: VMAF (Aaron et al. 2015), PSNR, and SSIM.

Table 18 Impact of normalization on accuracy using test data over 3 train/test split

Normalized data	BodyMTS accuracy
True	0.84
False	0.88

Table 19 Accuracy and video quality metrics score on test data over a single train/test split for different values of CRF

CRF	Average VMAF score	Average PSNR	Average SSIM	BodyMTS accuracy
16	97.03	48.14	0.998	0.83
22	95.54	44.51	0.996	0.83
28	91.21	39.85	0.990	0.82
34	87.76	36.52	0.979	0.75
40	67.94	33.09	0.954	0.69

FFmpeg has been utilized to calculate these metrics for different CRF values. We report the average metric score over all the clips for each value of CRF. Table 19 shows the average score over all the clips and the accuracy obtained using a particular value of CRF.

Results and discussion We observe that the VMAF score is more useful than other scores for estimating the quality of videos. Higher VMAF, indicates a better quality of videos. There is a big drop in the average VMAF score by changing the CRF values. Based on these results the threshold of VMAF can be set at around 90 which can be used to exclude those videos whose VMAF score is less than 90.

References

- Aaron A, Li Z, Manohara M, Lin JY, Wu ECH, Kuo CCJ (2015) Challenges in cloud based ingest and encoding for high quality streaming media. In: 2015 IEEE international conference on image processing (ICIP), pp 1732–1736. <https://doi.org/10.1109/ICIP.2015.7351097>
- Adnan NMN, Ab Patar MNA, Lee H, Yamamoto SI, Jong-Young L, Mahmud J (2018) Biomechanical analysis using kinovea for sports application, vol 342, no 1, p 012097
- Ahmadi A, Mitchell E, Destelle F, Gowing M, O'Connor NE, Richter C, Moran K (2014) Automatic activity classification and movement assessment during a sports training session using wearable inertial sensors. In: 2014 11th international conference on wearable and implantable body sensor networks. IEEE, pp 98–103
- Andriluka M, Roth S, Schiele B (2009) Pictorial structures revisited: people detection and articulated pose estimation. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 1014–1021
- Argent R, Slevin P, Bevilacqua A, Neligan M, Daly A, Caulfield B (2018) Clinician perceptions of a prototype wearable exercise biofeedback system for orthopaedic rehabilitation: a qualitative exploration. *BMJ Open*. <https://doi.org/10.1136/bmjopen-2018-026326>
- Argent R, Slevin P, Bevilacqua A, Neligan M, Daly A, Caulfield B (2019) Wearable sensor-based exercise biofeedback for orthopaedic rehabilitation: a mixed methods user evaluation of a prototype system. *Sensors*. <https://doi.org/10.3390/s19020432>
- Azulay A, Weiss Y (2019) Why do deep convolutional networks generalize so poorly to small image transformations? *J Mach Learn Res* 20:184:1-184:25
- Baechle TR, Earle RW (2008) Essentials of strength training and conditioning. Human Kinetics, Champaign

- Bagnall AJ, Dau HA, Lines J, Flynn M, Large J, Bostrom A, Southam P, Keogh EJ (2018) The UEA multivariate time series classification archive, 2018. CoRR abs/1811.00075 [arXiv:1811.00075](https://arxiv.org/abs/1811.00075)
- Brennan L, Kessie T, Caulfield B (2020) Patient experiences of rehabilitation and the potential for an mhealth system with biofeedback after breast cancer surgery: Qualitative study. *JMIR Mhealth Uhealth* 8(7):e19721
- Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans Pattern Anal Mach Intell*
- Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017. IEEE Computer Society, pp 4724–4733. <https://doi.org/10.1109/CVPR.2017.502>
- Carreira J, Noland E, Banki-Horvath A, Hillier C, Zisserman A (2018) A short note about kinetics-600. CoRR abs/1808.01340 [arXiv:1808.01340](https://arxiv.org/abs/1808.01340)
- Choutas V, Weinzaepfel P, Revaud J, Schmid C (2018) Potion: pose motion representation for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
- Chu WCC, Shih C, Chou WY, Ahamed SI, Hsiung PA (2019) Artificial intelligence of things in sports science: weight training as an example. *Computer* 52(11):52–61
- Dajime PF, Smith H, Zhang Y (2020) Automated classification of movement quality using the microsoft kinect v2 sensor. *Comput Biol Med* 125:104021
- Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: Leonardis A, Bischof H, Pinz A (eds) Computer vision—ECCV 2006, 9th European conference on computer vision, Graz, Austria, May 7–13, 2006, proceedings, part II, lecture notes in computer science, vol 3952. Springer, pp 428–441. https://doi.org/10.1007/11744047_33
- Dantone M, Gall J, Leistner C, Gool LV (2013) Human pose estimation using body parts dependent joint regressors. In: Proceedings of the IEEE conference on computer vision and pattern recognition
- Decroos T, Schütte K, Beéck TOD, Vanwansseele B, Davis J (2018) AMIE: automatic monitoring of indoor exercises. In: Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, proceedings, part III. Springer. https://doi.org/10.1007/978-3-030-10997-4_26
- Dempster A, Petitjean F, Webb GI (2019a) Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. [arXiv:1910.13051](https://arxiv.org/abs/1910.13051)
- Dempster A, Petitjean F, Webb GI (2019b) Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. [arXiv preprint arXiv:1910.13051](https://arxiv.org/abs/1910.13051)
- Dempster A, Petitjean F, Webb GI (2020) Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Min Knowl Discov* 34(5):1454–1495. <https://doi.org/10.1007/s10618-020-00701-z>
- Dempster A, Schmidt DF, Webb GI (2021) Minirocket: a very fast (almost) deterministic transform for time series classification. *KDD21 abs/2012.08791* [arXiv:2012.08791](https://arxiv.org/abs/2012.08791)
- Dhariyal B, Nguyen TL, Gsponer S, Ifrim G (2020) An examination of the state-of-the-art for multivariate time series classification. In: Workshop on large scale industrial time series analysis, ICDM 2020
- Dhariyal B, Le Nguyen T, Ifrim G (2021) Fast channel selection for scalable multivariate time series classification. In: ECMLPKDD
- Espinosa HG, Lee J, James DA (2015) The inertial sensor: a base platform for wider adoption in sports science applications. *J Fit Res* 4(1)
- Fan H, Li Y, Xiong B, Lo WY, Feichtenhofer C (2020) Pyslowfast. <https://github.com/facebookresearch/slowfast>
- Fang HS, Xie S, Tai YW, Lu C (2017) RMPE: regional multi-person pose estimation. In: ICCV
- Faro A, Rui P (2016) Use of open-source technology to teach biomechanics. *EDUCAȚIE FIZICĂ ȘI SPORT* p 18
- Fathallah Elalem S (2016) Evaluation of hammer throw technique for faculty of physical education students using dartfish technology. *J Appl Sports Sci* 6(2):80–87
- Fawaz HI, Forestier G, Weber J, Idoumghar L, Muller PA (2019) Deep learning for time series classification: a review. *Data Min Knowl Discov* 33(4):917–963. <https://doi.org/10.1007/s10618-019-00619-1>
- Feichtenhofer C (2020) X3D: expanding architectures for efficient video recognition. CoRR abs/2004.04730. [arXiv:2004.04730](https://arxiv.org/abs/2004.04730)
- Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2, 2019. IEEE, pp 6201–6210. <https://doi.org/10.1109/ICCV.2019.00630>

- Giggins OM, Caulfield B (2015) Proposed design approach for an interactive feedback technology support in rehabilitation. Association for Computing Machinery, New York, NY, USA, REHAB '15. <https://doi.org/10.1145/2838944.2838953>
- Girshick RB, Donahue J, Darrell T, Malik J (2013) Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR abs/1311.2524 [arXiv:1311.2524](https://arxiv.org/abs/1311.2524)
- Gkioxari G, Arbelaez P, Bourdev LD, Malik J (2013) Articulated pose estimation using discriminative armlnet classifiers. In: Proceedings of the IEEE conference on computer vision and pattern recognition
- He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 2961–2969
- Hinojosa C, Niebles JC, Arguello H (2021) Learning privacy-preserving optics for human pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2573–2582
- Huang S, Gong M, Tao D (2017) A coarse-fine network for keypoint localization. In: Proceedings of the IEEE international conference on computer vision
- Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B (2016) Deepcruc: a deeper, stronger, and faster multi-person pose estimation model
- Ji S, Xu W, Yang M, Yu K (2010) 3d convolutional neural networks for human action recognition. In: Fürnkranz J, Joachims T (eds) Proceedings of the 27th international conference on machine learning (ICML-10), June 21–24, 2010, Haifa, Israel. Omnipress, pp 495–502. <https://icml.cc/Conferences/2010/papers/100.pdf>
- Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The kinetics human action video dataset. CoRR abs/1705.06950 [arXiv:1705.06950](https://arxiv.org/abs/1705.06950)
- Krizhevsky A, Sutskever I, Hinton GE (2012a) Imagenet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States, pp 1106–1114
- Krizhevsky A, Sutskever I, Hinton GE (2012b) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
- Kwon H, Tong C, Haresamudram H, Gao Y, Abowd GD, Lane ND, Plötz T (2020) Imutube: automatic extraction of virtual on-body accelerometry from video for human activity recognition. Proc ACM Interact Mob Wearable Ubiquitous Technol 4(3):87. <https://doi.org/10.1145/3411841>
- Löning M, Bagnall A, Ganesh S, Kazakov V, Lines J, Király FJ (2019) skitime: a unified interface for machine learning with time series. In: Workshop on systems for ML at NeurIPS 2019
- Moral-Muñoz JA, Esteban-Moreno B, Arroyo-Morales M, Cobo MJ, Herrera-Viedma E (2015) Agreement between face-to-face and free software video analysis for assessing hamstring flexibility in adolescents. J Strength Cond Res 29(9):2661–2665
- Nakano N, Sakura T, Ueda K, Omura L, Kimura A, Iino Y, Fukushima S, Yoshioka S (2020) Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. Front Sports Act Living. <https://doi.org/10.3389/fspor.2020.00050>
- Nakano N, Sakura T, Ueda K, Omura L, Kimura A, Iino Y, Fukushima S, Yoshioka S (2020) Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. Front Sports Act Living 2:50
- Newell A, Huang Z, Deng J (2017) Associative embedding: end-to-end learning for joint detection and grouping
- O'Reilly M, Whelan D, Chaniyalidis C, Friel N, Delahunt E, Ward T, Caulfield B (2015) Evaluating squat performance with a single inertial measurement unit. In: 2015 IEEE 12th international conference on wearable and implantable body sensor networks (BSN). IEEE, pp 1–6
- O'Reilly MA, Whelan DF, Ward TE, Delahunt E, Caulfield BM (2017) Classification of deadlift biomechanics with wearable inertial measurement units. J Biomech 58:155–161
- Osokin D (2018) Real-time 2d multi-person pose estimation on cpu: lightweight openpose. arXiv preprint [arXiv:1811.12004](https://arxiv.org/abs/1811.12004)
- O'Reilly M, Caulfield B, Ward T, Johnston W, Doherty C (2018) Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review. Sports Med 48(5):1221–1246
- Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson J, Bregler C, Murphy KP (2017) Towards accurate multi-person pose estimation in the wild


- Pasos-Ruiz A, Flynn M, Bagnall A (2020) Benchmarking multivariate time series classification algorithms. [arxiv:2007.13156](https://arxiv.org/abs/2007.13156)
- Peng X, Wang L, Wang X, Qiao Y (2014) Bag of visual words and fusion methods for action recognition: comprehensive study and good practice. *CoRR abs/1405.4506*
- Pishchulin L, Insafutdinov E, Tang S, Andres B, Andriluka M, Gehler PV, Schiele B (2015) Deepcut: joint subset partition and labeling for multi person pose estimation
- Puig-Diví A, Escalona-Marfil C, Padullés-Riu JM, Busquets A, Padullés-Chando X, Marcos-Ruiz D (2019) Validity and reliability of the kinovea program in obtaining angles and distances using coordinates in 4 perspectives. *PLoS one* 14(6):e0216448
- Ressman J, Rasmussen-Barr E, Grooten WJA (2020) Reliability and validity of a novel kinect-based software program for measuring a single leg squat. *BMC Sports Sci Med Rehabil* 12:1–12
- Richter C, O'Reilly M, Delahunt E (2021) Machine learning in sports science: challenges and opportunities. *Sports Biomech*. <https://doi.org/10.1080/14763141.2021.1910334>
- Ruiz AP, Flynn M, Large J, Middlehurst M, Bagnall AJ (2021) The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Min Knowl Discov* 35(2):401–449. <https://doi.org/10.1007/s10618-020-00727-3>
- Sánchez J, Perronnin F, Mensink T, Verbeek JJ (2013) Image classification with the fisher vector: theory and practice. *Int J Comput Vis* 105(3):222–245. <https://doi.org/10.1007/s11263-013-0636-x>
- Sapp B, Taskar B (2013) MODEC: multimodal decomposable models for human pose estimation
- Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A (2016) Hollywood in homes: crowd-sourcing data collection for activity understanding. In: Leibe B, Matas J, Sebe N, Welling M (eds) *Computer vision—ECCV 2016—14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, proceedings, part I, lecture notes in computer science, vol 9905*. Springer, pp 510–526. https://doi.org/10.1007/978-3-319-46448-0_31
- Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ (eds) *Advances in neural information processing systems 27: annual conference on neural information processing systems 2014, December 8–13 2014, Montreal, Quebec, Canada*, pp 568–576
- Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings*. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556)
- Singh A, Le BT, Le Nguyen T, Whelan D, O'Reilly M, Caulfield B, Ifrim G (2020) Interpretable classification of human exercise videos through pose estimation and multivariate time series analysis. In: *5th international workshop on health intelligence at AAAI*. https://doi.org/10.1007/978-3-030-93080-6_14
- Slembrouck M, Luong H, Gerlo J, Schütte K, Van Cauwelaert D, De Clercq D, Vanwanseele B, Veelaert P, Philips W (2020) Multiview 3d markerless human pose estimation from openpose skeletons. In: Blanc-Talon J, Delmas P, Philips W, Popescu D, Scheunders P (eds) *Advanced concepts for intelligent vision systems*
- Soomro K, Zamir AR, Shah M (2012) UCF101: a dataset of 101 human actions classes from videos in the wild. *CoRR abs/1212.0402* [arXiv:1212.0402](https://arxiv.org/abs/1212.0402)
- Stamm O, Heimann-Steinert A (2020) Accuracy of monocular two-dimensional pose estimation compared with a reference standard for kinematic multiview analysis: Validation study. *JMIR Mhealth Uhealth* 8(12):e19608
- Tomar S (2006) Converting video formats with ffmpeg. *Linux J* 2006(146):10
- Tran D, Bourdev LD, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: *2015 IEEE international conference on computer vision, ICCV 2015, Santiago, Chile, December 7–13, 2015*. IEEE Computer Society, pp 4489–4497. <https://doi.org/10.1109/ICCV.2015.510>
- Trejo EW, Yuan P (2018) Recognition of yoga poses through an interactive system with kinect device. In: *2018 2nd international conference on robotics and automation sciences (ICRAS)*, pp 1–5. <https://doi.org/10.1109/ICRAS.2018.8443267>
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat FY, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van

- Mulbregt P, SciPy 10 Contributors, (2020) SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat Methods* 17:261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- Wang H, Schmid C (2013) Action recognition with improved trajectories. In: IEEE international conference on computer vision, ICCV 2013, Sydney, Australia, December 1–8, 2013. IEEE Computer Society, pp 3551–3558. <https://doi.org/10.1109/ICCV.2013.441>
- Wang X, Girshick RB, Gupta A, He K (2017) Non-local neural networks. *CoRR* abs/1711.07971 [arXiv:1711.07971](https://arxiv.org/abs/1711.07971)
- Whelan D, O'Reilly M, Huang B, Giggins O, Kechadi T, Caulfield B (2016) Leveraging imu data for accurate exercise performance classification and musculoskeletal injury risk screening. In: 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 659–662
- Whelan D, Delahunt E, O'Reilly M, Hernandez B, Caulfield B (2019) Determining interrater and intrarater levels of agreement in students and clinicians when visually evaluating movement proficiency during screening assessments. *Phys Ther* 99(4):478–486
- Zerpa C, Lees C, Patel P, Pryszucha E, Patel P (2015) The use of microsoft kinect for human movement analysis. *Int J Sports Sci* 5(4):120–127

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Ashish Singh¹  · Antonio Bevilacqua¹ · Thach Le Nguyen¹ · Feiyan Hu² · Kevin McGuinness² · Martin O'Reilly³ · Darragh Whelan³ · Brian Caulfield⁴ · Georgiana Ifrim¹

Antonio Bevilacqua
antonio.bevilacqua@insight-centre.org

Thach Le Nguyen
thach.lenguyen@insight-centre.org

Feiyan Hu
feiyan.hu@insight-centre.org

Kevin McGuinness
kevin.mcguinness@insight-centre.org

Martin O'Reilly
martin@outputsports.com

Darragh Whelan
darragh@outputsports.com

Brian Caulfield
b.caulfield@insight-centre.org

Georgiana Ifrim
georgiana.ifrim@insight-centre.org

¹ Insight Centre for Data Analytics, School of Computer Science, University College Dublin, Dublin, Ireland

-
- ² Insight Centre for Data Analytics, School of Electronic Engineering, Dublin City University, Dublin, Ireland
 - ³ Output Sports Limited, NovaUCD, Dublin, Ireland
 - ⁴ Insight Centre for Data Analytics, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland