An Examination of Wearable Sensors and Video Data Capture for Human Exercise Classification

Ashish Singh[1(✉)], Antonio Bevilacqua[1], Timilehin B. Aderinola[1],
Thach Le Nguyen[1], Darragh Whelan[2], Martin O'Reilly[2], Brian Caulfield[1],
and Georgiana Ifrim[1(✉)]

[1] Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland
{ashish.singh,antonio.bevilacqua,timi.aderinola,thach.lenguyen,
b.caulfield,georgiana.ifrim}@insight-centre.org
[2] Output Sports Limited, NovaUCD, Dublin, Ireland
{darragh,martin}@ouputsports.com

Abstract. Wearable sensors such as Inertial Measurement Units (IMUs) are often used to assess the performance of human exercise. Common approaches use handcrafted features based on domain expertise or automatically extracted features using time series analysis. Multiple sensors are required to achieve high classification accuracy, which is not very practical. These sensors require calibration and synchronization and may lead to discomfort over longer time periods. Recent work utilizing computer vision techniques has shown similar performance using video, without the need for manual feature engineering, and avoiding some pitfalls such as sensor calibration and placement on the body. In this paper, we compare the performance of IMUs to a video-based approach for human exercise classification on two real-world datasets consisting of Military Press and Rowing exercises. We compare the performance using a single camera that captures video in the frontal view versus using 5 IMUs placed on different parts of the body. We observe that an approach based on a single camera can outperform a single IMU by 10 percentage points on average. Additionally, a minimum of 3 IMUs are required to outperform a single camera. We observe that working with the raw data using multivariate time series classifiers outperforms traditional approaches based on handcrafted or automatically extracted features. Finally, we show that an ensemble model combining the data from a single camera with a single IMU outperforms either data modality. Our work opens up new and more realistic avenues for this application, where a video captured using a readily available smartphone camera, combined with a single sensor, can be used for effective human exercise classification.

Keywords: Exercise Classification · Inertial Sensors · Video · Time Series Classification · Real-World Datasets

1 Introduction

Recent years have seen an accelerated use of machine learning solutions to assess the performance of athletes. New technologies allow easier data capture and efficient machine learning techniques enable effective measurement and feedback. In this paper, we focus on the application of human exercise classification where the task is to differentiate normal and abnormal executions for strength and conditioning (S&C) exercises. S&C exercises are widely used for rehabilitation, performance assessment, injury screening and resistance training in order to improve the performance of athletes [18,19]. Approaches to data capture are either sensor-based or video-based. For sensor-based approaches, sensors such as Inertial Measurement Units (IMUs) are worn by participants [18,19]. For video, a participant's motion is captured using 3D motion capture [15], depth-capture based systems [31], or 2D video recordings using cameras [22,25]. The data obtained from these sources is processed and classified using machine learning models. Classification methods based on sensor data are popular in the literature and real-world applications, and yet, video-based approaches are gaining popularity [25,26] as they show potential for providing high classification accuracy and overcoming common issues of inertial sensors. Sensors require fitting on different parts of the body and the number of sensors to be worn depends upon the context of the exercise. For instance, the Military Press exercise requires at least 3 IMUs for optimal performance. Despite their popularity, sensors may cause discomfort, thereby hindering the movement of participants. In addition, using multiple sensors leads to overheads such as synchronization, calibration and orientation.
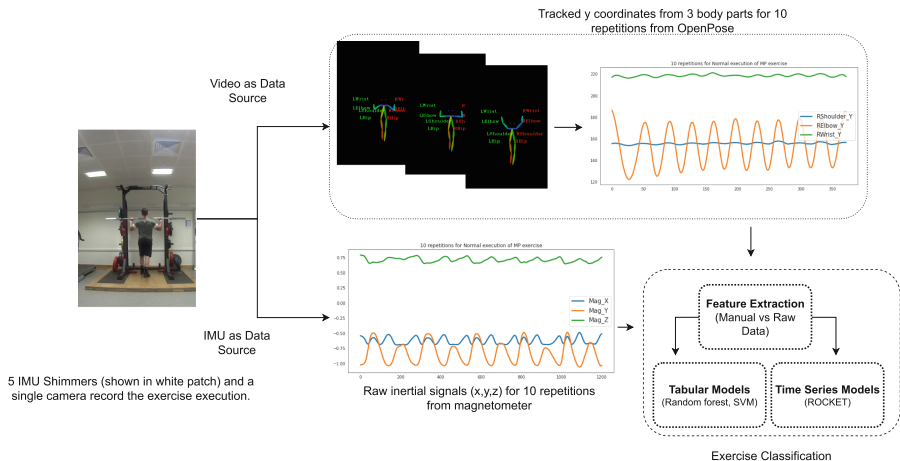

Fig. 1. Comparison of video (top) and sensors (bottom) to classify human exercise movement. The upper box presents the process of obtaining multivariate data from video (only 3 out of 25 body parts shown). The bottom box shows the raw Y-signals from a single IMU placed on the participant's body (only 3 signals shown here).

Recent advances in computer vision have enabled the usage of 2D videos for human exercise classification. Past work explored posture detection [22] and the application of human exercise classification using pose estimation. Our previous work [25] proposed a novel method named BodyMTS to classify human exercises using video, human pose estimation and multivariate time series classification. There is less work comparing sensors with video in real-world applications. In this paper, we compare the performance of a sensor-based approach utilizing 5 IMUs with that of video from a single front-facing camera, on the same set of 54 participants, on two real-world datasets consisting of Military Press (MP) and Rowing exercises. These are important S&C exercises and are widely used for injury risk screening and rehabilitation [30]. Incorrect executions may lead to musculoskeletal injuries and undermine the performance of athletes [1]. Hence, correct detection of abnormal movements is crucial to avoid injuries and maximize performance.

The main requirements for an effective human exercise classification application are [25]: accurate monitoring of body parts movement, correct classification of deviations from normal movements, timely feedback to end users, simple data capture using available smartphones and coverage of a wide range of S&C exercises. Previous work [29] has shown that this task is difficult and has poor intra and inter-rater accuracy in user studies with domain experts, with Kappa scores for inter-rater agreement between 0.18–0.53, and intra-rater between 0.38–0.62. Through discussions with domain experts, we established that an effective application should achieve a minimum accuracy of 80% to be useful for end users.

Existing methods using IMUs involve pre-processing the raw data, creating handcrafted features [18,20], and applying classical machine learning algorithms. Handcrafted feature extraction is often tedious and time-consuming, requires access to domain knowledge and is prone to cherry-pick features that only work for a specific set of exercises. Deep learning methods [17] overcome this issue by automatically constructing features during training, but still require expertise in deep learning architectures along with hardware resources such as GPUs. Hence, we take two approaches to feature extraction: (1) using lightweight packages such as catch22 [13] and tsfresh [4] to automate the feature extraction from raw signals and (2) using the raw time series data with time series classifiers, which implicitly construct features inside the algorithm. For videos, we first extract multivariate data using human pose estimation with OpenPose [3] to obtain (X,Y) location coordinates of key body parts over all the frames of a video. Figure 1 shows data captured with IMUs and video for the Military Press exercise. The top part shows the Y-signal for 3 body parts for a total of 10 repetitions, while the bottom part shows the X, Y, and Z signals of the magnetometer from an IMU worn on the right arm for the same set of 10 repetitions. **Our main contributions are:**

– We compare 3 strategies for creating features from IMU data for human exercise classification. We observe that directly classifying the raw signals using multivariate time series classifiers outperforms the approach based on handcrafted features by a margin of 10 and 4 percentage points in accuracy for MP and Rowing respectively. Automatic feature extraction shows better performance than handcrafted features.

- We compare the performance of IMU and video for human exercise classification. We observe that a single video-based approach outperforms a single IMU-based approach by a margin of 5 percentage points accuracy for MP and 15 percentage points for Rowing. Additionally, we observe that a minimum of 3 IMU devices are needed to outperform a single video for both exercises.
- We propose an ensemble model that combines the data modalities from IMU and video, which outperforms either approach by a minimum of 2 percentage points accuracy for both MP and Rowing. This leads to an accuracy of 93% for MP and 87% for Rowing, using only a single IMU and a reduced-size video. We discuss reasons why combining video and sensor data is beneficial, in particular, the 2D video provides positional information, while the sensor provides information on orientation and depth of movement.
- To support this paper we have made all our code and data available [1].

The rest of the paper is organized as follows. Section 2 presents an overview of related work, Sect. 3 describes the data collection procedure, Sect. 4 describes the data analysis and methodology for classification and Sect. 5 presents the classification results using IMUs and video. Section 6 concludes and outlines directions for future work and Sect. A discusses ethical implications of this work.

2   Related Work

This section describes the purpose of S&C exercises and provides an overview of sensor-based and video-based data capture approaches.

2.1   S&C Exercise Classification

S&C exercises aim at improving the performance of human participants in terms of strength, speed and agility, and they can be captured using sensor-based or video-based techniques.

Wearable sensor-based approaches involve fitting Inertial Measurement Units (IMUs) [18,19] on different parts of the body. This is followed by creating hand-crafted features which are used in conjunction with a classical machine learning model. Deep learning methods attempt to automate the process of feature extraction. CNN models work by stacking IMU signals into an image [17], whereas [28] uses an attention mechanism to identify the important parts in a signal. Using IMUs has its own limitations. First, the number of inertial sensors required and their positions can vary from exercise to exercise [18,20,30]. Furthermore, sensors require calibration and synchronization and may also hinder the movement of the body and cause discomfort when used over longer time periods [11,30].

Video-based systems can be categorized into 3 types: 3D motion capture, depth camera-based and 2D video camera. Though they are accurate, 3D motion capture systems are expensive and require complex setups. In addition, fitting multiple markers on the body may hinder the normal movement of the body [18].

---

[1] https://github.com/mlgig/Video_vs_Shimmer_ECML_2023.

Microsoft Kinect is commonly used for depth camera-based systems [5,23,31]. These systems are less accurate and are affected by poor lighting, occlusion, and clothing, and require high maintenance [18]. The third subcategory uses video-based devices such as DSLR or smartphone cameras. Works based on video rely on human pose estimation to track different body parts [16,25,26] and have shown 2D videos to be a potential alternative to IMU sensors. The video-based analysis also includes commercial software such as Dartfish [9] by providing the option to analyze motion at a very low frame rate. However, these are less accurate and require fitting body markers of a different colour to the background.

2.2   Multivariate Time Series Classification (MTSC)

In multivariate time series classification tasks, the data is ordered and each sample has more than one dimension. We focus on recent linear classifiers and deep learning methods, which have been shown to achieve high accuracy with minimal run-time and memory requirements [24,27].

Linear Classifiers. ROCKET [6] is a state-of-the-art algorithm for MTSC in terms of accuracy and scalability. Two more extensions named MiniROCKET [7] and MultiROCKET [27], have further improved this method. These classifiers work by using a large number of random convolutional kernels which capture different characteristics of a signal and hence do not require learning the kernel weights as opposed to deep learning methods. These features are then classified using a linear classifier such as Logistic or Ridge Regression.

Deep Learning Classifiers. Deep learning architectures based on Fully Convolutional Networks (FCN) and Resnet [10,24] have shown competitive performance for MTSC, without suffering from high time and memory complexity.

3   Data Collection

Participants. 54 healthy volunteers (32 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 were asked to complete multiple repetitions of the two exercises in this study; the Military Press and Rowing exercises. In each case, the exercises were performed under 'normal' and 'induced' conditions. In the 'normal' condition the exercise was performed with the correct biomechanical form and in the 'induced' condition the exercise was purposefully performed with pre-determined deviations from the normal form, assessed and confirmed in real-time by the movement scientist. Please refer to these sources [25,26] for additional information on the experiment protocol.

The data was collected using two video cameras and 5 Shimmer IMUs placed on 5 different parts of the body. Two cameras (30 frames/sec with 720p resolution) were set up in front and to the side of the participants. In this work, we only use the video recordings from the front view camera which is a more

common use case. The 5 IMUs with settings: sampling frequency of 51.2 Hz, tri-axial accelerometer(± 2 g), gyroscope ($\pm 500°$/s) and magnetometer (± 1.9 Ga) [20] were fitted on the participants at the following five locations: Left Wrist (LW), Right Wrist (RW), Left Arm (LA), Right Arm (RA) and Back. The orientation and locations of all the IMUs were consistent for all the participants.

Exercise Technique and Deviations. The induced forms were further subcategorized depending on the exercise.

3.1 Exercise Classes for Military Press (MP)

Normal (N): This class refers to the correct execution, involving lifting the bar from shoulder level to above the head, fully extending the arms, and returning it back to shoulder level with no arch in the back. The bar must be stable and parallel to the ground throughout the execution. **Asymmetrical (A):** The bar is lopsided and asymmetrical. **Reduced Range (R):** The bar is not brought down completely to the shoulder level. **Arch (Arch):** The participant arches their back during execution. Figure 2 shows these deviations using a single frame.
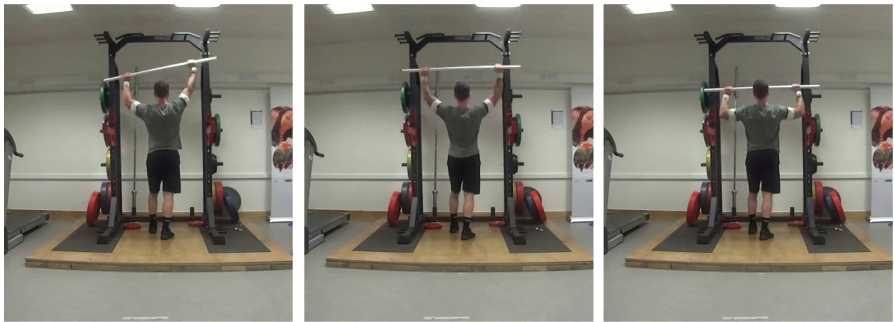

Fig. 2. Single frames from the Military Press exercise, depicting the induced deviations for class A, Arch and R (left to right).

Fig. 3. Single frames for the Rowing exercise, depicting the induced deviations for class A, Ext, R and RB (left to right).

3.2   Exercise Classes for Rowing

Normal (N): This class refers to the correct execution, where the participant begins by positioning themselves correctly, bending knees and leaning forward from the waist. The execution starts by lifting the bar with fully extended arms until it touches the sternum and bringing it back to the starting position. The bar must be stable and parallel to the ground and the back should be straight. **Asymmetrical (A):** The bar is lopsided and asymmetrical. **Reduced Range (R):** The bar is not brought up completely until it touches the sternum. **Ext:** The participant moves his/her back during execution. **RB:** The participant executes with a rounded back. Figure 3 shows these deviations by depicting a single frame.

4   Data Analysis and Methods

This section presents the data pre-processing, features extraction and classification models. We present the feature extraction for IMU data, followed by feature extraction for video. We also provide a description of the train/test splits for IMUs and video data.

4.1   IMU Data

We discuss three strategies to create features from IMU data. First, we directly use the raw signal as a time series. Second, we use existing approaches to create handcrafted features. Third, we use dedicated packages to automatically extract features. Features extraction is performed after segmenting the full signal to obtain individual repetitions.

Raw Signal as Multivariate Time Series. The raw signal from IMU records data for 10 repetitions. Hence, we segment the time series to obtain signals for individual repetitions. The Y signal of the magnetometer from the IMU placed on the right arm is utilized to segment the signals. The time series obtained after this step has variable length since the time taken to complete each repetition differs from participant to participant. Further, current implementations of selected time series classifiers cannot handle variable-length time series and therefore all time series are re-sampled to a length of 161 (length of the longest time series). This does not impact the performance as shown in the supplementary material. Every single repetition constitutes a single sample for train/test data. The final data D has a shape of $D \in \mathbb{R}^{N \times 45 \times 161}$, where N indicates the total samples. Each sample denoted by x_i in the data has a dimension of $x_i \in \mathbb{R}^{45 \times 161}$, where 45 denotes the total number of time series (5 IMUs x 9 signals) and 161 is the length of each time series.

Handcrafted Features. Each of the 5 IMUs outputs 9 signals (X, Y, Z) for each of the accelerometer, magnetometer and gyroscope. We follow the procedure as described in [20] to create handcrafted features. Additionally, 5 signals were created for each IMU: pitch, roll, yaw signal and vector magnitude of accelerometer and gyroscope, giving a total of 70 signals $(5 \times (9 + 5))$. For each repetition signal, 18 handcrafted features that capture time and frequency domain characteristics were created. Hence, we obtain the final data $D \in \mathbb{R}^{N \times 1260}$, where N is the total samples and 1260 represents the features extracted from 70 signals with 18 features each for both MP and Rowing.

Auto Extracted Features. We use packages catch22 [13] and tsfresh [4] to perform automatic feature extraction from a single repetition signal. These packages calculate a wide range of pre-defined metrics in order to capture the diverse characteristics of a signal. They are straightforward to use and avoid the need for domain knowledge and signal processing techniques. Catch22 captures 22 features for each of the 45 signals (5 IMUs x 9 signals) giving a total of 990 tabular features for MP and Rowing in the final dataset $D \in \mathbb{R}^{N \times 990}$, where N indicates the total samples. Similarly, tsfresh captures a large number of time series characteristics by creating a large number of features. The final dataset D has a shape of $D \in \mathbb{R}^{N \times 15000}$ and $D \in \mathbb{R}^{N \times 16000}$, for MP and Rowing respectively. Both manual and automatic feature extraction are performed on the normalized time series, as we observed that normalizing the time series leads to an increase in accuracy.

4.2 Video Data

We follow the methodology presented in our previous work [25] to classify human exercise from videos. OpenPose is used for human pose estimation to track the key body parts, followed by a multivariate time series classifier. Each video consists of a sequence of frames where each frame is considered a time step. Each frame is fed to OpenPose which outputs coordinates (X, Y) for 25 body parts. We only use the 8 upper body parts most relevant to the target exercises but also conduct experiments with the full 25 body parts. The time series obtained from a single body part is denoted by $b^n = [(X, Y)^1, (X, Y)^2, (X, Y)^3, ... (X, Y)^T]$ where n indicates the nth body part and T is the length of the video clip.

Multivariate Time Series Data. Since each video records 10 repetitions for each exercise execution, segmentation is necessary in order to obtain single repetitions. Each repetition forms a single time series sample for training and evaluating a classifier. We use peak detection to segment the time series as mentioned in our previous work [25]. Similarly to the IMU case, every time series obtained after this step has a variable length and therefore is re-sampled to a length of 161. The final data is denoted by $D \in \mathbb{R}^{N \times 16 \times 161}$, where N indicates the total samples. Each sample denoted by x_i has a dimension of $x_i \in \mathbb{R}^{16 \times 161}$, where 16 indicates X and Y coordinates for 8 body parts and 161 is the length of each time series.

Auto Extracted Features. We use catch22 [13] and tsfresh [4] to perform automatic feature extraction from each single repetition signal.

4.3   Train/Test Splits

We use 3 train/test splits in the ratio of 70/30 on the full data set to obtain train and test data for both IMUs and video. Each split is done based on the unique participant IDs to avoid leaking information into the test data. Train data is further split in the ratio of 85/15 to create validation data to fine-tune the hyperparameters. The validation data is merged back into the train data before the final classification. The data is balanced across all the classes. Table 1 shows the number of samples across all classes for a single train/test split for MP and Rowing respectively.

Table 1. Samples per class in train/test dataset for a single 70/30 split for MP (left) and Rowing (right) for both IMU and video.

Class	Train	Test	Total
N	370	150	520
A	340	150	490
R	366	155	521
Arch	350	140	490
Total	**1426**	**595**	**2021**

Class	Train	Test	Total
N	360	160	520
A	362	150	512
Ext	340	130	470
R	380	150	530
RB	361	140	501
Total	**1803**	**730**	**2533**

4.4   Classification Models

We use tabular machine learning models to work with handcrafted and automated features. Informed by previous literature on feature extraction for IMU data [18,20], we focus on Logistic Regression, Ridge Regression, Naive Bayes, Random Forest and SVM as classifiers for tabular data. We select ROCKET, MultiROCKET and deep learning models FCN and Resnet as recent accurate and fast multivariate time series classifiers [2].

5   Empirical Evaluation

We present results on IMU data, video data and combinations using ensembles. We report average accuracy over 3 train/test splits for all the results. We use the *sklearn* library [21] to classify tabular data and *sktime* [12] to classify time series data. All the experiments are performed using Python on an Ubuntu 18.04 system (16GB RAM, Intel i7-4790 CPU @ 3.60GHz). The Supplementary Material[2]

---

[2] https://github.com/mlgig/Video_vs_Shimmer_ECML_2023/blob/master/
Supplementary_material.pdf.

presents further detailed results on leave-one-participant-out cross-validation, demographic results, execution time, as well as the impact of normalization and re-sampling length on the classification accuracy.

5.1   Accuracy Using IMUs

We present the classification results using 3 different strategies for creating features from IMU data. For tabular features, we perform feature selection to reduce overfitting and execution time. We use Lasso Regression *(C=0.01)* with L1 penalty for feature selection, where C is the regularization parameter. Logistic Regression achieves the best performance followed by Ridge Regression and SVM. These results suggest that linear classifiers are best suited for this problem. Hence we only present results using Logistic Regression here. We tune hyperparameters, particularly regularization parameter C of Logistic Regression using cross validation. We observed that Logistic Regression (LR) with *C=0.01* achieves the highest accuracy (Table 3 presents results with Logistic Regression).

Table 2. Average accuracy on test data over 3 splits for selected multivariate time series classifiers using IMU raw data as time series.

Classifier	Acc MP	Acc Rowing
FCN	0.86	0.77
ResNet	0.87	0.74
ROCKET	**0.91**	**0.80**
MultiROCKET	**0.91**	**0.81**

Table 2 presents the results using raw data and multivariate time series classifiers. ROCKET achieves the best performance with MultiROCKET having similar accuracy for this problem. ROCKET has the added benefit that it can also work with unnormalised data and it is faster during training and prediction, so we select this classifier for the rest of the analysis. We analyse the average accuracy using all 5 IMUs as well as combinations of IMUs using raw time series with ROCKET as classifier. The goal is to select the minimum number of IMUs needed to achieve the best performance for MP and Rowing. Table 3 presents the average accuracy over 3 splits obtained using all IMUs whereas Table 4 presents the average accuracy using different combinations of IMUs.

Results and Discussion: From Table 3 we observe that using raw data with ROCKET achieves the highest accuracy when compared to the approaches based on handcrafted and automated feature extraction. We tune hyperparameters of ROCKET using the validation data, particularly the *number-of-kernels* and observe no impact on the accuracy. The normalization flag is set to True here as turning it off leads to a 4 percentage points drop in the accuracy. ROCKET can easily be run on a single CPU machine without the need for much engineering

Table 3. Average accuracy obtained on 5 IMUs data by using three feature selection strategies. Logistic Regression (LR) is used for tabular data, whereas ROCKET is used for time series classification.

Feature Type	Acc MP	Acc Rowing
Tabular		
Handcrafted	0.80	0.76
Automated (catch22)	0.84	0.75
Automated (tsfresh)	0.88	0.80
Raw Signals		
Time series	**0.91**	**0.80**

Table 4. Average accuracy obtained using the different placement of IMUs over three train/test splits using raw data as time series with ROCKET as classifier.

Placement of IMU	Acc MP	Acc Rowing
5 IMUs	**0.91**	**0.80**
RightWrist	0.83	0.68
LeftWrist	0.84	0.70
RightArm	0.77	0.65
LeftArm	0.76	0.66
Back	0.71	0.71
LeftWrist + RightWrist	0.88	0.75
LeftWrist + RightWrist + Back	**0.91**	**0.80**
LeftArm + RightArm	0.82	0.70
LeftArm + RightArm + Back	0.86	0.78

effort (only 2 parameters to tune) and dedicated hardware. It is much faster than using tsfresh or catch22 for feature extraction followed by classification. Table 4 presents the accuracy using different combinations of IMUs placed on different parts of the body. Accuracy is lowest when using only a single sensor. Accuracy starts to increase as more IMUs are included, for both MP and Rowing. We observe that placing 1 IMU on each wrist and 1 at the back achieved the same accuracy as using all 5 IMUs. The accuracy jumps from 0.83 to 0.88 moving from one IMU placed on the right wrist to two IMUs placed on both wrists and finally jumps to 0.91 when adding one more IMU at the back for MP. Similar behaviour is observed for Rowing. This suggests that 3 IMUs are sufficient for these exercises.

5.2   Accuracy Using Video

Here we present the results of classification using video as the data source. We report the average accuracy over 3 train/test splits for MP and Rowing. We also

Table 5. Average accuracy obtained by ROCKET using video as data source for MP and Rowing over three train/test splits.

Feature Type	Acc MP	Acc Rowing
Tabular		
Automated (catch22)	0.69	0.70
Automated (tsfresh)	0.77	0.77
Raw Signals		
25 body parts	0.82	0.79
8 body parts	**0.88**	**0.83**
Elbow Pair [8]	0.83	0.82

present results using tabular classifiers with automated features for comparison with the IMU based approach. For the raw data approach, we study the accuracy when involving different body parts, e.g., all 25, the 8 upper body parts suggested by domain experts and results using automated channel selection technique [8]. The normalization flag is set to False here as turning it on leads to a 4 percentage points drop in accuracy. This is in contrast to the setting configured for IMUs. We tune hyperparameters of ROCKET, particularly the *number-of-kernels* and observe no impact on the accuracy. Table 5 presents the average accuracy using these different approaches for classifying MP and Rowing exercises.

Results and Discussion: From Table 5 we observe that the average accuracy achieved using raw time series is highest when using the 8 body parts suggested by domain experts. Using automated features does not seem to work very well, in this case, achieving accuracy below 80% for both exercises. Moreover, using channel selection techniques leads to an improvement by 1 and 3 percentage points in accuracy versus using the full 25 body parts.

5.3 IMU Versus Video

We compare IMU and video data for human exercise classification, using the raw data approach for both IMU and video as it achieves the best performance. We report the accuracy, the execution time and the storage space required.

Table 6 presents the results for both MP and Rowing exercises. We observe that a minimum of 3 IMUs are required to achieve a higher accuracy than a single video. A single video outperforms a single IMU for both exercises by a minimum of 5 percentage points. Table 7 reports the real train/test time for both approaches. This time includes time taken for data pre-processing and to train/test the model. It also includes time to run pose estimation in case of video. The IMUs approach takes the least amount of time to train/test as compared to the video-based approach. For video, OpenPose extracts the multivariate time series data. The total duration of all videos is 1 h 38 min for MP, whereas Open-Pose took 1 h 12 min thus OpenPose can run faster than real-time, which is

Table 6. Comparison of accuracy obtained using IMUs and video for MP and Rowing.

Data Source	Acc MP	Acc Rowing
Placement of IMUs		
3 IMUs (Wrists and Back)	0.91	0.80
1 IMU (LeftWrist)	0.84	0.70
Video		
25 body parts	0.82	0.79
8 body parts	0.88	0.83
Ensemble: video and IMUs		
Video (8 body parts) + 3 IMUs	**0.93**	**0.88**
Video (8 body parts) + 1 IMU LeftWrist	**0.93**	0.87

important for getting fast predictions. Table 8 presents the storage consumption for both approaches. We note savings in terms of storage space: 5 IMUs require 6 times more space than the time series obtained from videos. Even after selecting the minimum number of sensors which is 3 in both exercises, the storage consumption is more than 200 MB which is also higher as compared to using time series from video. Our previous work in [25] explored the impact of video quality such as resolution and bit rate on classification accuracy and demonstrated how much video quality can be degraded without having a significant impact on the accuracy, whilst saving storage space and processing power.

5.4   Combining IMU and Video

We create an ensemble model by combining individual models trained independently on IMU and Video. For IMUs, we take the 3 sensors that achieved the highest accuracy. When video is combined with just a single sensor, we take the IMU placed on the left wrist, as it had the highest accuracy among single sensors and it is the most common location for people to wear their smartwatch. Probabilities are combined by averaging and the class with the highest average probability is predicted for a sample during test time. Table 6 presents a comparison of different approaches, using ROCKET as a multivariate time series classifier. From Table 6, we observe that an ensemble model achieves the best average accuracy when compared to using any number of IMUs and a single video-based approach. The accuracy for MP jumps by 2 percentage points when transitioning from 5 IMUs to an ensemble approach, and by 5 percentage points when moving from a single video to an ensemble. Similar results are observed for Rowing. These results suggest that combining IMU and video modalities enhances the performance of exercise classification. Combining video and IMU data sources, with video providing 2D location coordinates for key anatomical landmarks and IMUs capturing acceleration and orientation of the body parts, results in improved classification accuracy, as shown in this investigation

Table 7. Average train/test time (minutes) obtained using IMUs and video as data sources for MP over three train/test splits. We also report the average test (i.e., prediction) time over a clip of 10 repetitions.

Data Type	Training Time (minutes)	Test Time/ **Test time per clip of 10 reps**
Sensor		
3 IMUs (Wrists and Back)	8	6/0.10
Video		
8 body parts	52	22/0.37
Ensemble: video and IMUs		
Video (8 body parts) + 5 IMUs	60	29/0.50
Video (8 body parts) + 1 IMU	58	27/0.46

Table 8. Storage consumption using raw videos, IMUs and video as time series for MP and Rowing exercises for the 54 participants in our study.

Data Size (MB)	MP	Rowing
5 IMUs	640	591
Raw Videos (720p)	813	1012
Videos as Time Series	97	114

(see supplementary material). This finding is consistent with previous work in [14] that highlights the complementary nature of video and IMUs in enhancing human pose estimation quality, while in this work we see a similar benefit for human exercise classification.

6   Conclusion

We presented a comparison of IMU and video-based approaches for human exercise classification on two real-world S&C exercises (Military Press and Rowing) involving 54 participants. We compared different feature-creation strategies for classification. The results show that an automated feature extraction approach outperforms classification that is based on manually created features. Additionally, directly using the raw time series data with multivariate time series classifiers achieves the best performance for both IMU and video. While comparing IMU and video-based approaches, we observed that using a single video significantly outperforms the accuracy obtained using a single IMU. Moreover, the minimum number of IMUs required is not known in advance, for instance, 3 IMUs are required for MP to reach a reasonable accuracy. Next, we compared the performance of an ensemble method combining both IMU and video with the standalone approaches. We showed that an ensemble approach outperforms either data modality deployed in isolation. The accuracy achieved was 93% and

88% for MP and Rowing respectively. The criteria to select sensors or videos will ultimately depend on the goal of the end user. For instance: the choice between video and IMUs will depend on a combination of factors such as convenience and levels of accuracy required for the specific application context.

We acknowledge the fact that the scenario that was tested in this research does not accurately reflect real-world conditions. This does mean that we are exposed to the risk that the induced deviation performances could be exaggerated, and therefore not reflective of the often very minor deviations that can be observed in the real-world setting. However, we would argue that performing exercises under induced deviation conditions, if done appropriately, is a very necessary first step towards validating these exercise classification strategies in this field. It would not be prudent to assume that this model could be generalised to operate to the same level in real-world conditions. Having said that, the use of conditioned datasets is a necessary first step in this kind of application and provides the proof of concept evidence necessary to move onto the real-world setting.

A   Ethical Implications

Using videos for human exercise classification raises ethical implications that need to be mitigated, prompting a discussion of potential ethical implications.

Data Collection. Participants in this study provided written consent and the Human Research Ethics Committee of the university approved this study. All experiments were conducted under the supervision of an expert physiotherapist. The potential implications, in this case, can arise when the language used for the consent form may not be native to all the participants. In our case, the organizing authority or professional who was carrying out the data collection made sure that all the participants have well understood the consent form and the use of this data in the future.

Privacy and Confidentiality. This study uses videos which record participants executing exercises. This poses obvious privacy challenges. A first step is to blur the video to protect the participant's identity. This work utilizes human pose estimation to extract time series from video, thereby avoiding the need to directly use the original video. By working with the extracted time series, it largely safeguards the privacy and confidentiality of the participants.

Diversity of Representation. The participants considered in this study fall into the age group of 20 to 46. Hence the results presented here may not generalise for other age groups. Therefore the final use case will depend on the specific target users, such as athletes competing in the Olympic games versus individuals

with less intensive training goals. While there were slightly more male participants than female participants, it does not impact the conclusions drawn in this work, as analysed in the supplementary material. However, this requires further exploration to avoid any biases in the conclusion. Future studies should aim for equal representation among participants in terms of age, sex, gender, race etc., from the start of the study.

Transparency and Feedback. The prediction of the model in this case outputs whether the execution of the exercise was correct or incorrect. Deep learning-based models and other posthoc explanation methods support saliency maps which can be used to highlight the discriminative regions of the data that can be mapped back to the original video thus providing more information about the model decision to the participant.

The above list is not exhaustive and other inherent biases may appear because of the chosen model and the way the data has been collected.

References

1. Baechle, T.R., Earle, R.W.: Essentials of strength training and conditioning. Human kinetics, Champaign, IL (2008)
2. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. In: Data Mining and Knowledge Discovery, pp. 1–55 (2016). https://doi.org/10.1007/s10618-016-0483-9
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: real-time multi-person 2d pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
4. Christ, M., Braun, N., Neuffer, J., Kempa-Liehr, A.W.: Time series feature extraction on basis of scalable hypothesis tests (tsfresh - a python package). Neurocomputing **307**, 72–77 (2018). https://doi.org/10.1016/j.neucom.2018.03.067
5. Decroos, T., Schütte, K., De Beéck, T.O., Vanwanseele, B., Davis, J.: AMIE: automatic monitoring of indoor exercises. In: Brefeld, U., et al. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11053, pp. 424–439. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10997-4_26
6. Dempster, A., Petitjean, F., Webb, G.I.: Rocket: exceptionally fast and accurate time series classification using random convolutional kernels. Data Min. Knowl. Discov. **34**(5), 1454–1495 (2020), https://doi.org/10.1007/s10618-020-00701-z
7. Dempster, A., Schmidt, D.F., Webb, G.I.: Minirocket: a very fast (almost) deterministic transform for time series classification. In: KDD 2021 arXiv: 2012.08791 (2021)
8. Dhariyal, B., Le Nguyen, T., Ifrim, G.: Scalable classifier-agnostic channel selection for multivariate time series classification. Data Min. Knowl. Discov. **37**(2), 1010–1054 (2023). https://doi.org/10.1007/s10618-022-00909-1
9. Faro, A., Rui, P.: Use of open-source technology to teach biomechanics. Educaţie Fizică şi Sport, p. 18 (2016)
10. Fawaz, H.I., Forestier, G., Weber, J., Idoumghar, L., Muller, P.A.: Deep learning for time series classification: a review. Data Min. Knowl. Discov. **33**(4), 917–963 (2019). https://doi.org/10.1007/s10618-019-00619-1

11. Kwon, H., et al.: Imutube: automatic extraction of virtual on-body accelerometry from video for human activity recognition. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**(3), 87 (2020). https://doi.org/10.1145/3411841
12. Löning, M., Bagnall, A., Ganesh, S., Kazakov, V., Lines, J., Király, F.J.: sktime: a Unified Interface for Machine Learning with Time Series. In: Workshop on Systems for ML at NeurIPS 2019 (2019)
13. Lubba, C.H., Sethi, S.S., Knaute, P., Schultz, S.R., Fulcher, B.D., Jones, N.S.: catch22: Canonical time-series characteristics selected through highly comparative time-series analysis. bioRxiv (2019)
14. von Marcard, T., Pons-Moll, G., Rosenhahn, B.: Human pose estimation from video and imus. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 1533–1547 (2016)
15. Molías, L.M., Ranilla, J.M.C., Cervera, M.G.: Pre-service physical education teachers' self-management ability: a training experience in 3d simulation environments. Retos: nuevas tendencias en educación física, deporte y recreación (32), 30–34 (2017)
16. Nakano, N., et al.: Evaluation of 3d markerless motion capture accuracy using openpose with multiple video cameras. Front. Sports Active Living **2** (2020). https://doi.org/10.3389/fspor.2020.00050
17. Nutter, M., Crawford, C.H., Ortiz, J.: Design of novel deep learning models for real-time human activity recognition with mobile phones. In: 2018 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2018). https://doi.org/10.1109/IJCNN.2018.8489319
18. O'Reilly, M., Caulfield, B., Ward, T., Johnston, W., Doherty, C.: Wearable inertial sensor systems for lower limb exercise detection and evaluation: a systematic review. Sports Med. **48**(5), 1221–1246 (2018)
19. O'Reilly, M., et al.: Evaluating squat performance with a single inertial measurement unit. In: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 1–6. IEEE (2015)
20. O'Reilly, M.A., Whelan, D.F., Ward, T.E., Delahunt, E., Caulfield, B.M.: Classification of deadlift biomechanics with wearable inertial measurement units. J. Biomech. **58**, 155–161 (2017)
21. Pedregosa, F., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
22. Rahmadani, A., Bayu Dewantara, B.S., Sari, D.M.: Human pose estimation for fitness exercise movement correction. In: 2022 International Electronics Symposium (IES), pp. 484–490 (2022). https://doi.org/10.1109/IES55876.2022.9888451
23. Ressman, J., Rasmussen-Barr, E., Grooten, W.J.A.: Reliability and validity of a novel kinect-based software program for measuring a single leg squat. BMC Sports Sci. Med. Rehabil. **12**, 1–12 (2020)
24. Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M., Bagnall, A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. In: Data Mining and Knowledge Discovery, pp. 1–49 (2020)
25. Singh, A., et al.: Fast and robust video-based exercise classification via body pose tracking and scalable multivariate time series classifiers. In: Data Mining and Knowledge Discovery (Dec 2022). https://doi.org/10.1007/s10618-022-00895-4
26. Singh, A., et al.: Interpretable classification of human exercise videos through pose estimation and multivariate time series analysis. 5th International Workshop on Health Intelligence at AAAI (2020). https://doi.org/10.1007/978-3-030-93080-6_14

27. Tan, C.W., Dempster, A., Bergmeir, C., Webb, G.I.: MultiRocket: multiple pooling operators and transformations for fast and effective time series classification. arxiv:2102.00457 (2021)
28. Tao, W., Chen, H., Moniruzzaman, M., Leu, M.C., Yi, Z., Qin, R.: Attention-based sensor fusion for human activity recognition using imu signals. arXiv: 2112.11224 (2021)
29. Whelan, D., Delahunt, E., O'Reilly, M., Hernandez, B., Caulfield, B.: 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 (2019)
30. Whelan, D., O'Reilly, M., Huang, B., Giggins, O., Kechadi, T., Caulfield, B.: 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), pp. 659–662. IEEE (2016)
31. Zerpa, C., Lees, C., Patel, P., Pryzsucha, E., Patel, P.: The use of microsoft kinect for human movement analysis. Inter. J. Sports Sci. **5**(4), 120–127 (2015)