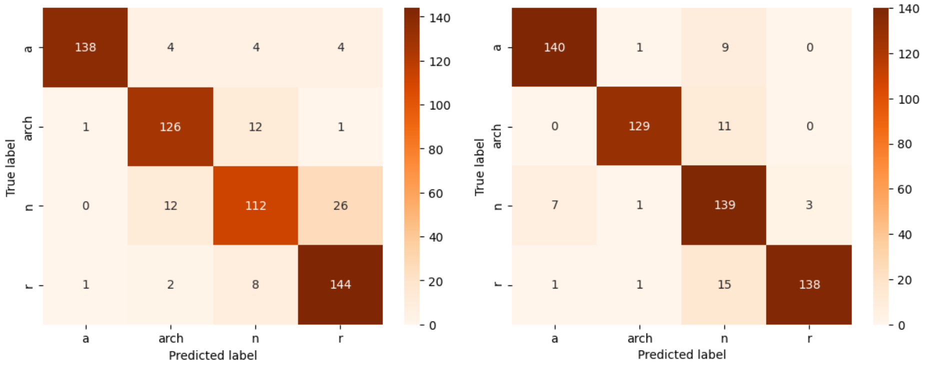
1 Supplementary Material

In this section, we present additional results from the comparison of IMUs and video-based approaches for human exercise classification. We show the confusion matrix for the MP exercise for a single split, leave-one-subject-out-cross-validation (LOSOCV) results, and the impact of normalization and re-sampling on the classifier performance. We further present the classification results based on demographic factors, execution time taken and explore the underlying mechanism behind the improved performance of the ensemble model.

1.1 Confusion matrix.

We show the confusion matrix for a single train/test split on the MP exercise to compare performance using IMUs and a video-based approach. The confusion matrix is obtained using all the 5 IMUs whereas for videos the confusion matrix is obtained using 8 upper body parts. The total instances for each class are the sum of all numbers from each column.

Fig. 1. Confusion matrix for Video (Left) and IMU (Right) for MP exercise over a single split.



From Figure 1, we observe that video-based classification can achieve similar performance in terms of precision and recall for each of the classes.

1.2 Cross Validation Results.

We present the leave-one-subject-out-cross-validation (LOSOCV) results for three different strategies used to create features IMU data: handcrafted features, automated features and using raw data directly. Logistic Regression is used to classify the features extracted using handcrafted and automated features, whereas ROCKET is used in the case of directly using raw data. In the case of video, we only present the LOSOCV results for 8 body parts using ROCKET.

Table 1. Leave-one-subject-out-cross-validation (LOSOCV) accuracy obtained using various feature creation strategies from IMUs and video using for MP and Rowing.

Data Approach	Accuracy on MP	Accuracy on Rowing
Sensor		
Handcrafted features (LR)	0.80	0.71
Auto features catch22 (LR)	0.82	0.73
Auto features using tsfresh (LR)	0.85	0.80
Raw time series 5 IMUs (ROCKET)	0.88	0.80
Video		
8 body parts (ROCKET)	0.88	0.83

From Table 1 we observe that using raw data directly from IMUs achieves a higher performance than other approaches based on creating handcrafted or automated features. Additionally, LOSOCV accuracy is better when using videos for classification for the Rowing exercise. These results indicate that raw data from IMUs can be utilized directly without the need for manual and automated feature engineering. Furthermore, video can also be used as an alternative source for data collection for human exercise classification. These results confirm the previous results shown in the main article.

1.3 Impact of normalization

We consider the impact of normalization on classification accuracy. We use ROCKET as a classifier for both IMU and video. The normalization flag is turned on by default in the case of ROCKET. We only present results for the MP exercise here.

Table 2. Impact of normalization on average accuracy using ROCKET as a classifier for MP exercise over 3 train/test splits.

Normalized data?	Accuracy MP	Accuracy Rowing
Sensor		
True (default in ROCKET)	0.91	0.80
False	0.87	0.80
Video		
True (default in ROCKET)	0.84	0.73
False	0.88	0.83

From the above results, it can be concluded that the video-based approach requires the normalization flag to be set to False. Setting the flag to True leads to a loss of magnitude information making it difficult for the classifier to capture distinguishing patterns for different classes. However, in the case of IMU, setting normalization to True helps in improving the accuracy. This is probably because it brings all values on the same scale for different numbers of participants where the time and variation to complete single repetition changes from participant to participant. However, this flag has no impact on the accuracy of Rowing. Hence, further investigation is required by including more exercises which is outside the scope of this work.

1.4 Impact of length re-sampling.

We consider the impact of length re-sampling on the accuracy. The work in [1] showed that accuracy remains unaffected for different values of re-sampling in the case of videos. Here we explore the impact of re-sampling when using raw data directly from IMUs with ROCKET as a classifier. The default value of re-sampled length is 161 in both cases.

Table 3. Impact of length re-sampling using raw data from IMUs as time series with ROCKET as a classifier for MP exercise over 3 train/test splits.

Re-sampled Length	**Accuracy on MP**
100	0.90
161	0.91
200	0.91
300	0.91
400	0.91

From the above results, it can be concluded that re-sampling the data to different lengths does not produce any significant impact on the accuracy. However, a higher value for length leads to higher consumption of storage space.

1.5 Impact of demographic features on classification

We analyse the impact of demographic factors such as sex, age and BMI (Body Mass Index) on the classification accuracy for both video and IMU-based approach. We only present results for the MP exercise here. We present the average accuracy over 3 train/test splits here. There are 7 male and 7 female participants in the testing data on average.

Table 4 shows the results for each gender, and Table 5 presents the accuracy for each age group. We segment the age by taking a step size of 10 and Table 6 shows the accuracy for each BMI group. The total samples shown in each table present the average number of samples over 3 splits in test data. In the case of gender, video performs better for the female gender than the IMU-based approach. It should be noted that there are more males than females overall. In the case of age, video performs better for the age group 40-50. In the case of BMI, the video-based approach performs better for Healthy weight participants

Table 4. Depicting the classification accuracy across each gender group over 3 train/test splits for MP.

Sex	Total Samples	Acc Video	Acc IMU
Female	310	0.87	0.83
Male	279	0.84	0.85

Table 5. Depicting the classification accuracy across each age group group over 3 train/test splits for MP.

Age Group	Total Samples	Acc Video	Acc IMU
20 - 30	423	0.85	0.82
30 - 40	125	0.88	0.88
40 - 50	41	0.96	0.84

Table 6. Depicting the classification accuracy across each BMI group over 3 train/test splits for MP.

Sex	Total Samples	Acc Video	Acc IMU
Healthy Weight	394	0.89	0.83
Obesity	13	0.95	0.97
Overweight	154	0.77	0.83
Underweight	27	0.86	0.87

and IMU perform better for Overweight participants. These results conclude that video can achieve similar performance as IMUs.

1.6 Train/test Time.

Here we present the time taken to train/test for each of the methods used to classify the IMU data. The raw data is used directly as time series using ROCKET as the multivariate time series classifier. The final data consists of signals from all 5 sensors. The train/test time includes the data-processing time. Table 7 shows the train/test time (in seconds) for all methods for the MP exercise. From these results, it is clear that the ROCKET takes the least amount of time to train and FCN takes the least amount of time to test. However, the combined train and test time of ROCKET is lower than FCN. Among the deep learning classifiers, FCN is the fastest probably because of a simpler architecture and fewer parameters as compared to Resnet.

1.7 Analyzing Ensemble Model.

We conducted further experiments in an attempt to understand the mechanism underlying the increased accuracy of the ensemble model. The ensemble model

Table 7. Train/test time (in seconds) averaged over 3 train/test splits for all the classifiers for the MP exercise.

Classifier	Train Time	Test Time
FCN	970	4
Resnet	1060	5
ROCKET	106	40
MultiROCKET	368	80

which is composed of two independent models trained separately on video and IMU data. Here IMU placed on the right wrist is considered. The final probability is calculated by averaging the probabilities from both the models and the class with the the maximum probability value is predicted.

Here we break down the ensemble model by combining the individual component such as the accelerometer (AR), magnetometer (MR) and gyroscope (GY). Separate models are trained on each of these components and their combinations and combined with a video-based model to create the final ensemble model. Table 8 presents the accuracy for these combinations on the MP exercise. We observed that signals from the accelerometer and gyroscope play a more significant role in the accuracy than from the magnetometer. This suggests that combining the orientation information from the IMUs with the positional information from videos plays an important role in improving accuracy. We also investigated weighted combinations of probabilities based on the model accuracy and observed similar results.

Table 8. Impact of each component and combinations of these on the average accuracy of ensemble model for the MP exercise over 3 train/test splits.

Combination	Accuracy on MP
Video + AR + MR + GY	0.93
Video + AR	0.91
Video + GY	0.90
Video + MR	0.89
Video + GY + AR	0.92
Video + GY + MR	0.91
Video + AR + MR	0.91

References

1. Singh, A., Bevilacqua, A., Nguyen, T.L., Hu, F., McGuinness, K., O'Reilly, M., Whelan, D., Caulfield, B., Ifrim, G.: Fast and robust video-based exercise clas-

sification via body pose tracking and scalable multivariate time series classifiers. Data Mining and Knowledge Discovery (Dec 2022). https://doi.org/10.1007/s10618-022-00895-4, https://doi.org/10.1007/s10618-022-00895-4