Movement recognition for physiotherapy

Salya Diallo, Shrinidhi Singaravelan, Fanny Ghez
Department of Computer Science, EPFL Lausanne, Switzerland
CS-433 : Machine Learning Project 2

***Abstract*—The aim of this paper is to showcase our project which consists of predicting human movements from multi-dimensional positional data. Using Machine Learning techniques seen in the course such as *Classification and Neural Networks*, our model demonstrates robust performance across different activities performed by different participants.**

I. Introduction

Motion recognition, which is the ability to understand and predict human movements, plays an important role in various domains such as healthcare or sports analytic. In our case, motion recognition is being used in physiotherapy. Our dataset has been sourced from the SMS lab specializing in physiotherapy, from Balgrist Campus a research campus located at the Balgrist University Hospital in Zurich.We shall present our exploration, where the goal was to develop one or more robust models allowing to accurately predict intricate human movements based on a big multi-dimensional dataset. Similarly as many articles like [1] that have different types of movements considered. We begin our paper with a description of the dataset, giving a deeper understanding to the reader. Then we shall provide a detailed overview of the pre-processing steps undertaken to ensure data quality and consistency and we go on explaining the intricacies of our model architecture, which consists of Classifications and Recurrent Neural Networks. We also studied classification using angles rather than positional values, which could be an interesting research subject for the future.

To evaluate the model's performance we adopted robust metrics such as accuracy and the F1 score, considering the multi-class nature of the recognition task. Moreover, we also used confusion matrices in order to visualize the model's strengths and areas of improvements across different activities.

Previous works in the domain employed several methods such as Random Forests [2], SVM classification techniques [3], [4] or complex algorithms such as Dynamic Time Warping (DTW) algorithms [5]. We have taken these projects into account, while reproducing some techniques and having new ideas of our own.

The outcomes of our project extend beyond model performance, as we delve into ethical consideration. We discuss implications related to privacy, bias and the responsible usage of motion recognition systems.

II. Materials and Methods

A. Dataset

1) Dataset Acquisition: The dataset consists of 2183485 data points and 104 features. There are 25 participants, where each can do 7 different exercises and the movement can be filmed by 4 different cameras, allowing to capture exercises in different ways. Hence, the different features are:

- Participant ID, from P04 to P28,
- Exercise (abduction, bird, bridge, knee, shoulder, squat or stretch),
- Set (tells us if there is an error in the movement),
- Time,
- Position of 33 different captors, given in meters, telling us the distance from the center of the hips of the corresponding participant to a certain body part, each on 3 different axis: x, y, or z.

The dataset provided by the laboratory is in 2 different files, an original one, and a relabeled one. It is the same as the original one, but with an additional column 'Label' containing the value '1' if the subject is in the exercise position, and '0' otherwise (T-pose, getting in position, etc...).

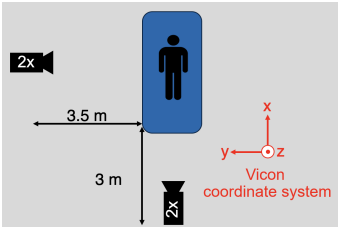


Fig. 1: Representation of how the data was collected: positions of the 4 cameras, the participant and the axis

B. Pre-processing

To ensure data quality, 27′187 rows were removed from the dataset, each containing more than 95% of missing values.

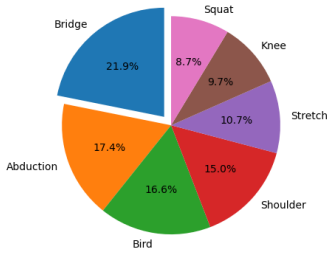


Fig. 2: Percentage of each exercise in the cleaned dataset

Normalization of the data was considered unnecessary due to the presence of a common reference frame. As all positional

data shares the same center point and frame, we saw that normalizing the data did not provide additional benefits for model training. We initially tried implementing MinMaxScaler from the *sklearn* library, but after testing all methods after normalization, the results were the same or even less accurate.

As per the data splitting, 4 distinct methods (M) were employed for splitting the data into train and test sets:

- **M1 :** we removed the time column and a split of 50% for testing and 50% for training was performed;
- **M2 :** the dataset was split based on participants and we kept the time column, with 12 participants reserved for testing and 13 for training;
- **M3 :** we took a percentage $p = 50\%$, and took p percent for the training set of each unique values of the feature '*Camera*', so that training take into consideration each cameras.
- **M4 :** we took a certain percentage $p = 50\%$, and took p percent for the training set of each unique values of the feature '*Set*', so that training take into consideration each errors.

For most of our models, we need the matrices X and our prediction target y to be defined. We consider y to be the encoded *Exercise* column and X to be the rest of our data set.

C. Exercise Classification Models

The main goal of our project is to classify each movement. As there are 99 different categorical variables, 33 per x, y, z, we choose to not remove any of these variables as they all are useful.

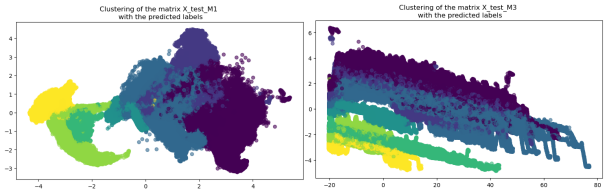
1) Clustering: We thought about using a k-clustering method, where k is known and is equal to 7. Nevertheless, clustering is useful for unsupervised learning whereas in this project we know what we want to predict and what each data point has to be. Hence, we did not pursue this method.

2) Classifiers: One way of approaching the problem is by classification. For the classification, we tried all train-test separation mentioned above, and kept the one that gave us the more accuracy, if there was a difference.

A question arose when analyzing the data set: *does the camera influence the prediction?* In order to answer that, we compared the accuracy obtained when classifying on the whole data set and on sub-data sets corresponding to each camera.

To predict the labels on the test set we used a classifier. We tried using SUPPORT VECTOR MACHINE (SVM) or RANDOM FOREST (RF) and kept the one with the best accuracy, if there was a difference. In order to find the best parameters for the diverse classifiers, we used a sub-sample of the training set and then used cross-validation combined with grid search. At first, we started with a sub-sample size of 10000, and then we tried to define the sub-sample size in function of the number of data points in the train set, to obtain more stable result. Hence we took a fraction of the number of data points of the train set, rounded to obtain an integer. We could have opted to only do cross-validation on the whole train matrices, but due to the size of them the running time was too high (with more than 500 minutes for RF for example), especially when considering the whole matrix X. By taking a fifth of the training set length, the running time is still really high for some splitting method (up to 300 minutes). If time permit, there is no issue but if the goal is for each cell to run in less than one hour, the size of the sample has to be reduced.

We note that Random Forest is a popular method in different articles like [4]. Recall that each tree in the forest is constructed using a random subset of the training data and a random subset of features, providing diversity among the trees.

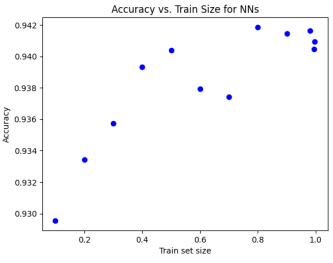


(a) M1 (b) M3

Fig. 3: Visualizations of the predicted clustering on the test set of the whole matrix X using splittings M1 and M3 using SVM classifier

For these classification methods, we remove the features *Participant*, *Set*, and *Camera* in order to only predict using the positional values, or time depending on the splitting method. In figure 3, we can see an example of the clustering we obtained using the SVM classifier and Principal Component Analysis (PCA), with two components, for the dimensional reduction. We used PCA only for visualization of the clusters since our data set has 99 dimensions, if we consider only positional values.

3) Neural Networks - Soft-max Entropy: This model, that we call a *SimpleClassifier*, features a straightforward architecture approach suitable for our motion classification.



(a) Train split accuracy

Train split	Accuracy
98.00%	94.16%
99.70%	94.09%
99.50%	94.05%
90.00%	94.15%
80.00%	94.19%
70.00%	93.74%
60.00%	93.79%
50.00%	94.04%
40.00%	93.93%
30.00%	93.57%
20.00%	93.34%
10.00%	92.95%

(b) Accuracy of train

Fig. 4: Train split for NN

The input layer, that has a size matching the dimension of the positional data, is connected to a hidden layer through a linear transformation. We then use Re-LU as an activation function which introduces non-linearity. And then the hidden layer connects to the output layer, which then provides logits

for each class, and the final prediction is obtained through a soft-max operation during evaluation, and 10 epochs are used. Moreover, the parameters were found doing a cross validation in order to be more precise. The percentage of train set we choose would be justified by Figure 4a, choosing percentages of train sets varying between 0.1 and 0.98. Thus, we have decided to take a train split of 50 %, due to the ratio between the high accuracy and computational cost.

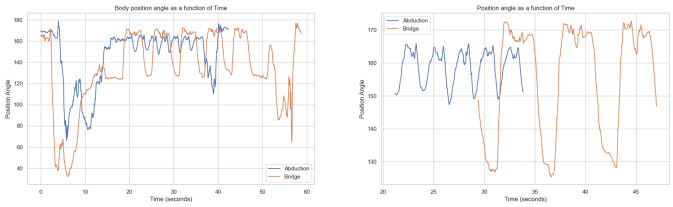
4) Neural Networks - Impact of error: In this approach, we created a Neural Network trained only on correct examples. Then, we tested it on the entire dataset to see how well it could predict outcomes when faced with new errors. This helps us understand how predictive algorithms might respond and how well they can detect small variations in a new user's exercise routine. The Neural Network in this method is similar to the Softmax Entropy approach, but the focus here is on figuring out how mistakes in exercise execution could affect motion recognition accuracy.

After trying different versions of Neural Network layering, the difference in accuracy between testing on 'Correct' sets only and over all the data set was never exceeding 5% which we consider to be acceptable in the context of motion recognition. The sampling frequency of the camera being :

$$f = \frac{1}{T_{sampling}} = \frac{1}{0.033s} = 30Hz$$

a 5% difference in accuracy would result in 45 wrongly predicted over 909 frames for a 30 sec exercise, corresponding to 1.5s error spread over the entire exercise.

5) Angular Analysis: Another method to address our dataset was to focus on calculated angles in the body position. We measure 6 different angles (3 on each side) : bending of both legs, both arms and bending of the overall body (angle drawn by shoulder - hip - knee) on both sides. But when plotting the dynamics of these angles over time, we noticed that the data provided would be largely biased by the early and late stages of each exercise as the subject first stand in a T pose, then claps their hand, get in position to begin the exercise, and leaves the position at the end of it. These stages can easily be visualized in figure 5.



(a) Full exercise (b) Active part of exercise

Fig. 5: Body angle in function of time

To avoid the angle corruption of these 2 stages, we use for this angular analysis the second dataset provided by the lab : relabeled dataset. By only keeping rows with '1'-label, we limit our analysis to the active period of each exercise (i.e. the period we are interested in predicting). Based on this time-series dataset, we can now train a SVM classifier to try and categorize exercises. This method resulted in a 94% accuracy which is quite good with regard to the small dimension (time column + 6 angles).

Another simpler approach was first to increase probability of different exercise based on angle ranges (min and max values of each angle reached for different exercises), but this method was not conclusive as ranges were pretty large and similar and almost half of the rows turned out to have angles in several ranges, resulting in undefined exercise, and 54% was the best accuracy achieved with this technique.

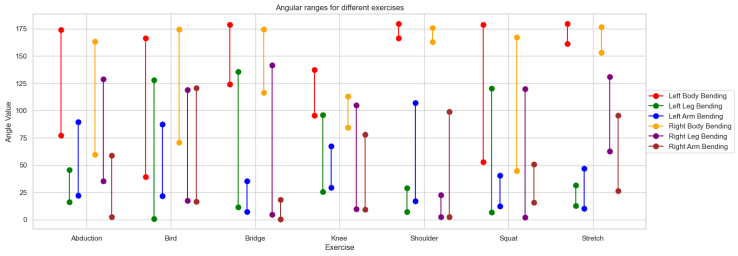


Fig. 6: Angle ranges for each exercise

In figure 6, we can see for example that a left arm bending (blue angle) of approx. 25° would cast a vote for 6 exercises out of 7, which would decrease accuracy of the predicted exercise.

III. Results

First of all, in order to visualize the different accuracies of each methods, we have chosen to look at them in confusion matrices (in fig. 7 for RF) using a similar justification as in paper [6].

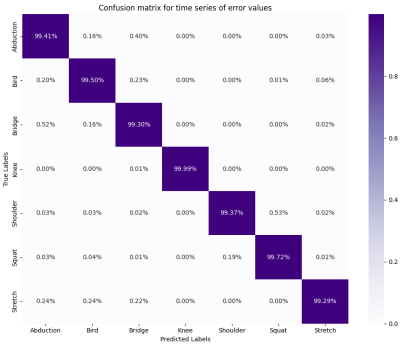


Fig. 7: RF based on method **M4**

Then we have an overall comparison of the different models according to the different methods described accordingly based on the article [6]: **M1, M2, M3** and **M4**.

We also wanted to see if cameras had an impact on prediction, to answer this question we look at table I.

Remark that when using a higher sampling size, for example a fifth of the number of data points in the training set, the accuracy for SVM is 100% for every method and matrices,

Accuracy		
Matrix	**SVM**	**RF**
X	92.06%	92.98%
X_Frontal_Top	93.62%	95.68%
X_Frontal_Low	93.39%	95.33%
X_Side_Top	93.40%	94.56%
X_Side_Low	93.50%	94.81%

TABLE I: Accuracy of the different implemented methods for the different matrices, using splitting M4 and a sample size of 10000 for the hyper-parameters tuning

but the running time is bigger (up to 5 hours). Hence there is an accuracy-running time trade-off to take into consideration, especially if the data set increases in size.

We can easily see in table I that in the two methods, separating the data set by cameras increases the accuracy. This comes from the fact that the perspective of the movement changes in function of the camera. Hence, using a frontal top camera is not the same as using a side low camera for example. Moreover, the article [7] did a similar study concluding that cameras did in fact have an impact on the results.

Now, in addition to this analysis, we delved into the study of which axis has the most impact on the prediction. Hence, using any of the classifier (since the result will proportionally be the same, so we chose SVM), we removed each axis and did three more classification on the whole matrix X, using splitting M3. We obtained the result that the z-axis had the less impact and the x-axis was the most important one. The x-axis being the depth of the movement, it makes sense that it is the most impactful one since most of the seven movements imply a modification in depth. Indeed, in figure 8 we can see that the three most recurrent exercises, that represent 55.9% of the data set's (cf. figure 2), vary a lot with time in function of the x-axis when we look at the knee of one of the participants. Remark that depending on the splitting method used, the y-

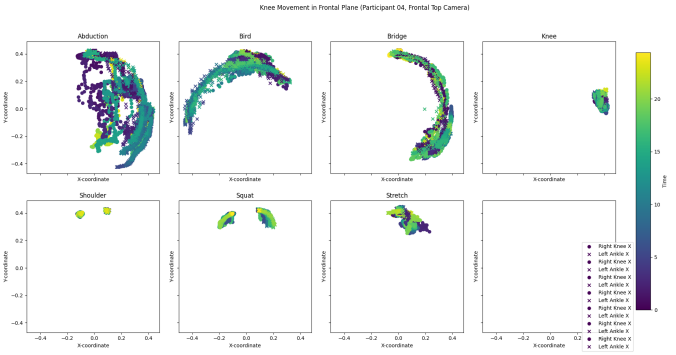


Fig. 8: Knee movement in the yx plane for each exercise in function of the time for participant P04 using the frontal top camera

axis may have more impact than removing the x-axis - this is the case for M2 for example. This may come from the fact that depending on the participant taken in the training set, the y-axis may have more importance than the x-axis.

Models	**M1**	**M2**	**M3**	**M4**	**Overall**	**F1 Score**
SVM	100%	100%	100%	100%	100%	1
RF	99.26%	93.47%	99.47%	99.46%	97.40%	0.976
NN	94.40%	93.25%	93.25%	95.40%	94.35%	0.951

TABLE II: Accuracy and F1 Score

IV. Discussion

Looking at table II, we can see that the accuracies are all above 93 %. The most accurate model is SVM, giving us a perfect accuracy of 100%. This may come from the fact that in our data set, some data points are not part of the movement but represent transitions between movements, creating noise. Or SVM can be more robust to noisy data than Random Forest, which could explain the better accuracy. After that we have *Random Forest*, with an overall accuracy of 97.4 %, with the highest accuracy being 99.47 % for **M3**. This may be justified by its big number of decision trees each trained on different data subsets and considering random features at each split, leading to a robust model that averages out errors and reduces overfitting. However, we must add that even though RF is a pretty accurate model, it is computationally very costly, leading to a trade-off between accuracy and computational cost.

Moreover, we can see that depending on the model, there are methods that have better accuracies than others. While exploring whether the cameras had an impact on the classification, we deduce that the answer is yes. Indeed, for the NN, we can see that the accuracy is 1% higher compared to the other ones, and similarly it is also a little higher for RF. Therefore, we could deduce that the choice of camera plays a significant role in our classification.

The confusion matrix shows that some exercises are better detected than other: for instance, the knee exercise has a near 100% accuracy, this is due to the fact that this exercise differs from the other ones and is easier to distinguish.

V. Conclusion

Many different methods were implemented during this project covering various visions of the recognition problem, arising diverse questions that we tried to answer. To tackle this questions, we decided to look at the problem from different perspective: treating positional values as time series or solely considering positions independently of time.

Nevertheless, to get a better accuracy higher running time is necessary and thus higher computational costs. One has to be careful to this trade-off and take it into consideration when using these models on other data sets.

But now, rather than using positional values, why not work with angles? That was one of our point of view, where we calculated 6 different angles. Removing the movements noise (movement that had nothing to do with the exercise), we obtained an accuracy of 94%. This could be a base model for future analysis, and a possible lead for broader application, for example a wider set of exercises.

ACKNOWLEDGEMENTS

We would like to thanks David Rode for the opportunity to work on this project and answering all our questions throughout the weeks. In addition, we thanks our teaching assistant Mohamed Hichem Hadhri for his help throughout the project and his helpful answers to our various questions.

VI. ETHICS

Impact of Training Bias on Healthy, Abled Subjects

A. Risk Description

The potential bias induced by training machine learning models on a dataset predominantly composed of young, healthy subjects without impairments poses a significant challenge to the robustness and applicability of the model across diverse demographics. Specifically, if the goal is to accurately predict exercises based on video or sensor data for new participants, relying solely on a dataset featuring young, paired, and healthy individuals is sub-optimal. The limitations become evident when faced with participants who deviate from the dataset norm, such as individuals with obesity, advanced age, or disabilities (e.g., amputees).

The inherent risk lies in the fact that movements may vary in terms of amplitude, timing, or other characteristics for individuals with different health profiles. Consequently, the model may fail to produce accurate predictions for this broader population, effectively rendering it tailored to a narrow segment of the demographic. This limitation is particularly problematic for individuals with specific health conditions or disabilities, as the model may not accommodate their unique movement patterns.

B. Risk Evaluation

The evaluation of this risk is rooted in the examination of the individuals constituting the training and testing datasets. Notably, all subjects within these datasets were characterized by perfect health and physical fitness. Consequently, when attempting to predict movements for individuals with diverse health backgrounds, the model may exhibit inaccuracies due to the absence of representation for those with specific health challenges, such as obesity, anorexia, paraplegia, or missing limbs.

C. Mitigating the Risk

While the existing dataset may be immutable, proactive measures can be implemented to overcome this bias and enhance the model's inclusivity. One recommendation is to advocate for Inclusive Data Collection practices. Collaborating with the data-providing laboratory to diversify the dataset by including individuals with varying age ranges and health conditions would contribute to a more accurate and representative model. This approach ensures that the model learns from a more comprehensive spectrum of movements, accommodating the unique characteristics of different user groups.

Furthermore, a practical solution involves providing information to users about potential limitations in detection accuracy for certain impairments. This proactive communication strategy can manage user expectations and inform them that precise detection may be challenging for specific impairments. It also serves as a mechanism to foster transparency and user awareness, contributing to a more responsible and user-centric deployment of the model.

REFERENCES

[1] E. M.-M. E. C. M. C. F. G.-D. Felix Escalona, "Eva: Evaluating at-home rehabilitation exercises using augmented reality and low cost sensors," *Virtual Reality manuscript*, 2020.

[2] T. M. M. H. Colin Arrowsmith, David Burns and C. Whyne, "Physiotherapy Exercise Classification with Single-Camera Pose Detection and Machine Learning," *sensors*, 29 December 2022.

[3] M. Y. A. B. Bappaditya Debnath, Mary O'Brien, "A review of computer vision-based approaches for physical rehabilitation and assessment," *Regular paper*, 19 June 2021.

[4] N. J. J. Q. B. S. D. B. M. E. H. Andrew Hua, Pratik Chaudhari, "Evaluation of machine learning models for classifying upper extremity exercises using inertial measurement unit-based kinematic data," *IEEE Journal of Biomedical and Health Informatics*, June 2020.

[5] C. H. Wentong Zhang, Caixia Su, "Rehabilitation Exercise Recognition and Evaluation based on smart Sensors with Deep Learning Framework," *IEEEAccess*, 2015.

[6] A. P. A. N. D. T. E. S. Artem Obukhov, Andrey Volkov and I. Fedorchuk, "Examination of the Accuracy of Movement Tracking Systems for Monitoring Exercise for Musculoskeletal Rehabilitation," *Sensors*, 24 Septembre 2023.

[7] I. J.-V. F. J. R.-R. M. G.-M. J. Z.-P. R. M.-S. R. M.-C. Rafael Aguilar-Ortega, Rafael Berral-Soler and M. J. Marín-Jiménez, "Uco physical rehabilitation: New dataset and study of human pose estimation methods on physical rehabilitation exercises," *Sensors*, 31 October 2023.

VII. ANNEX

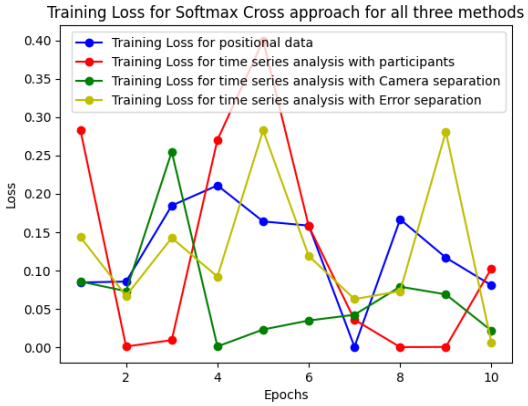


Fig. 9: Losses plot

Abbreviations

- NN : Neural Networks
- RF : Random Forests
- SVM : Support Vector Machine