Kinematic analysis of shoulder motion for diagnostic purposes

Arjun Sardjoe Missier, Lennart van Koppen, Brice Lang-Nguyen, Raphael Pickl, Hassan Ali, Eddie Versluis, Rachelle Kiepe, Dr. Tony Andrioli

Abstract

The purpose of this research is to classify patients with different degrees of severity of Rotator Cuff tears using data generated by a Flock of Birds system.

The data that is used during this research consists of movement of the bones around the shoulder. This is based on time series, where every frame contains the rotation angles of the bones. These recordings were processed, normalized and labeled by the LUMC.

Multiple configurations of the data are used to do multiclass classification with a logistic regression model. The best configuration resulted in an accuracy of 69.9%.

Results show that the model has difficulties in recognizing patient group (PG) 2. There are multiple assumptions on why this is the case, but none can be clarified indefinitely.

The results of multiple configurations show that data cleaning steps improve the model results, e.g. normalization, removing double exercises and only using 5 frames per exercise to create a input vector. Other cleaning steps decreased the accuracy of the model, e.g. removing idle, this can mean that the model can't recognize the complexity of the data without other variables that influences certain PGs more than others.

Introduction

The use of machine learning in the medical field started in the 1990's. With the increase of data and processing power in the 21st century, this became a more usable tool. Since 2010 there has been a sharp rise in companies, investing in machine learning and artificial intelligence with the purpose of supporting the medical field (D. Douglas Miller, 2018).

While most machine learning focuses on Radiology, Oncology, Cardiology and Pathology, the focus of Computer-Assisted-Diagnosis (CAD) lies in image processing and assessing bio markers. The Laboratory for Kinematics and Neuromechanics (LK&N) in the Leiden University Medical Center (LUMC) (Sylvia A. Stegeman, 2016) (C.G.M.Meskers, 1998) is searching for new tools to support doctors in diagnosing Musculoskeletal Disorders (MSD) that are not found within these areas. This research is a follow up on previous researches done by the students of the minor Applied Data Science at the Hague University of Applied Sciences and the LK&N (Kasper van der Hoofd, 2019) as to find new ways to diagnose Rotator Cuff tears using machine learning.

The purpose of this research is to redo the classification of patients with similar complaints and or degree of severity of Rotator Cuff tears using data generated by the Flock of Birds system. To do the classification a logistic regression model will be used, because previous group had good results with this model. The main goal thereby is to reduce the assumptions and come with clear results that a logistic

regression model is able to classify the patients.

This resulted to the following research question:

"To what extend and in what way, can different data science techniques be used on kinematic recordings to contribute to a valid and reliable diagnosis, made by a doctor, on Musculoskeletal Disorders."

Kinematic data used for our research contains recordings of different types of exercises from multiple patient. In most cases there were two or more recordings for each of the exercise types. Patients have been recorded in 4 different patient groups (**PG**) based on degree of severity of Rotator Cuff tears. All records have been labeled by LUMC physicians.

A full elaboration of these issues would be beyond the scope of this paper. This research will focus on the kinematic data gathered at The Laboratory for Kinematics and Neuromechanics (LK&N) from the LUMC where was collaborated with Dr. ir. J. H. de Groot, Head of Research and Development of LK&N (LUMC, 2020).

Techniques

Exercises are recorded using the Flock of Birds, this system measures positions through the means of multiple magnetic field sensors relative to a magnetic field generator (C.G.M. Meskers, 1999). Using boney landmarks as reference points to represent the positions and movements of bones in a cartesian space. The dataset consists of four PG's, containing between 30 and 40 patients. This data was recorded for four previous studies with four different protocols. Regarding to the protocols all PG's did the same five standard exercises, sometime twice, with an average length of >50 frames. Knowing the protocols gave certainty on what an exercise should look like. This data retrieved from the LUMC is the raw data, which is later transformed to Euclidian space.

Normalization is applied to the dataset for scaling all values within a set boundary. Fourier Transformation (Bevel, 2020) is in some cases applied to resample the number of datapoints of a signal.

Logistic regression (Hoffman, 2019) was used as a data science technique to make multiclass classification on three different PG. This model was chosen based on a research paper, that found Logistic regression to be able to classify PG's up to 97% accuracy (Kasper van der Hoofd, 2019). The issue with this research is that there were a lot of assumptions made on the labels of the columns and the exercises. Because of those assumption, the research is picked up again, but this time with labels for the columns and the protocols, which give the exercises their name.

To validate functionality of the trained model the following validation metrices are used:

- Accuracy
- Recall
- Precision
- Confusion matrix

These metrics will provide enough information to pick up on common machine learning mistakes like overfitting.

Methods

The raw data is in cartesian space, which contains information about the height of a patient and the size of its bones. To prevent any machine learning method from having bias towards the size of a patient's limbs, all bony landmarks have been converted to 3D Euler angles absolute to each other, thus not reliant on the other bones

Visualizing converted data showed several recordings contained more than one exercise. The files containing the recordings were manually inspected, and the start and end of individual exercises were noted. A script reading these notes extracted the recordings into separate files.

Exercise recordings could contain stationary data at the beginning and the end which does not contain useful information. Stationary data is part of a recording between the time a physician starts and/or stops the recording, and when the patient starts and/or stops moving. A script determined by looking at a plane, where a possible idle is most apparent, if a movement exceeds the average slope and corrects the recording.

General

Calculate the slope of the motion data
Filter the Slope
Split the exercise into equal parts
Calculate the mean of the for the first part
Calculate the mean of the last part

While *the script has not found an idle:* **do:**
loop through values of the filtered slope
if the value < mean
Save value into a list

Return: the index of the first value in the list

General form of the script Remove idle

Figure 1 remove_idle

For machine learning purposes all recordings have been normalized.

The data conversion process attempts to calculate the angle of an elbow. The output values of the script were found to be unrealistic because the values were too big to be considered elbow angles. Therefor elbow values (left & right) were considered unreliable thus not included in the input vector.

A TSNE (fig 2) (Maaten, 2014) plot pointed out the data conversion process could not smoothly convert PG4 and ended up producing questionable values. These values did not resemble the other PG's in any way. Therefor PG4 got dismissed as unreliable. This leaded to an investigation of the data conversion process to determine where it might have gone wrong. Looking through this process yielded the following points of concern:

- Difference in input metrices
- Incorrect calibration file
- Generated dummy data

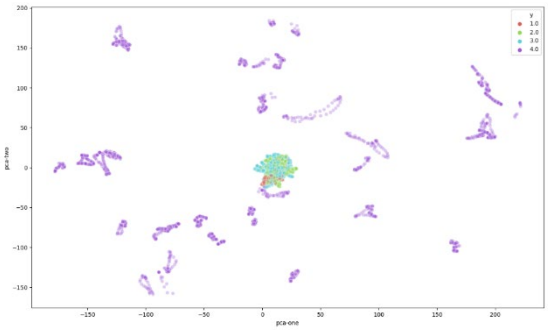

Figure 2 TSNE of PG 1-4

Improving the size of the data-set all available exercise types from a single patient are combined. To define this as an input vector the exercise is reshaped into a 1D vector (fig3). From this vector n frames are picked and evenly divided over the recording. Each of the picked frames from the combined exercises contains up to 8 bones which contain the three Euler Angles, these frames are appended to the final output vector.

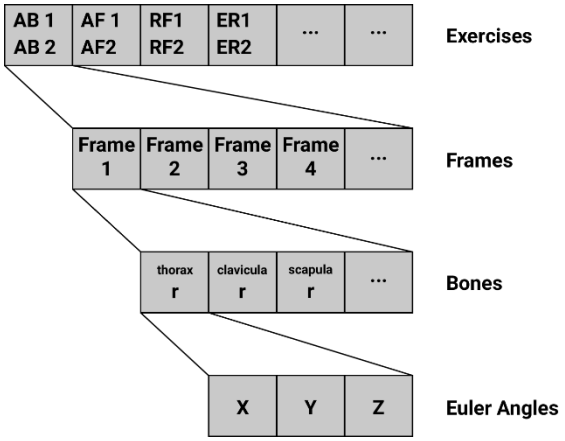

Figure 3 data structure

The selection of n frames over all the exercises reforms the input vectors over all patients to the same shape.

A cartesian product from the available exercises for a single patient is taken to find all possible unique combinations.

Every method of creating/composing an input vector for kinematic recordings has its own hyper parameters. E.g: the number of frames taken from a single exercise. To understand the importance of each individual parameter, more than 100 unique configurations of parameters were tested. Not all these configurations gave desired accuracy, that's why only the 5 best configurations get explained in this paper.

In order to reliantly compare different models, there always needs to be the same patients in the test and the train set. That's why no cross-validation was used, even though it is still questionable if there is enough data.

Result

The top 5 results are shown below in figure 4. The highest accuracy was 69.6%. The data used in this model was normalized with selection of columns which contains the kinematic data of the bones and axis as shown in figure (fig 8). Furthermore, the configuration also included normalization and it used 5 frames per exercise.

A multiclass confusion matrix is used to visualize the results of the different configurations are shown in figure (fig 5).

Accuracy	Remove Idle	Frame generator	Column Index	N frames	Normalizing	Resample	default
0.6955	False	False	0	5	True	False	True
0.6912	False	False	0	15	True	False	True
0.6907	False	False	1	5	True	False	True
0.6876	False	False	0	5	False	False	True
0.6860	True	False	1	15	True	False	True

Figure 4 best five results

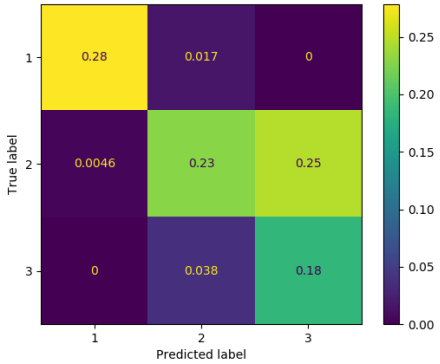

Figure 5 Confusion Matrix

The resulted precision and recall per group classification are as followed (fig 6+7).

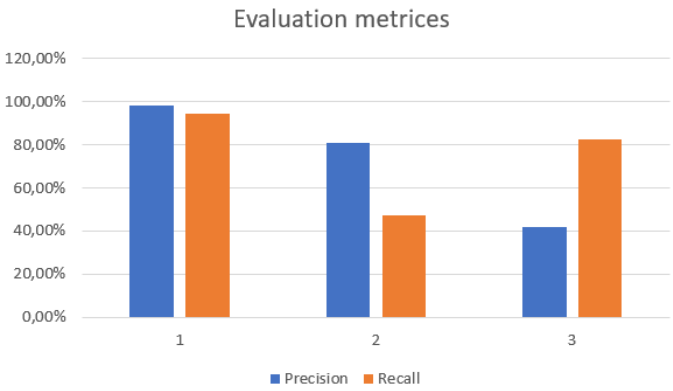

Figure 6 best result

Patient groups	Precision	Recall
1	98,30%	94,30%
2	80,70%	47,50%
3	41,90%	82,60%

Figure 7 best result

Bone	**Left**	**Right**
Clavicula	X	X
Scapula	X	X
Humerus	X	X
Thorax	X	X
Clavicula	X	X

Figure 8 bones for best config

Discussion

The results show that the model had difficulties in recognizing PG 2. It lies closely to PG 3 and this leads to the precision being high but recall being low for PG 2, while influences precision and recall of PG 3. The reason for this could be one of the following:

1. The computer model that calculates the small characteristic differences between PG's is not advanced enough to differentiate them. Resulting into similar classification for PG's that with similar protocol and diagnosis. To argue that the better and closer the protocols are, a more complex computer model can possibly differentiate between the possible MSD's characteristics instead of relying on the characteristics of the protocol which are possibly influenced by a physician. PG 2 is close to patient group 3 possibly because of the same protocol that has been used for to record individual exercises. In comparison to PG 1 these PG fit closely together, that makes the simple computer model perfect in defining in between PG 1 against PG 2 and 3. Thus leading to a bias in leading to PG 2 and 3.

2. It could also be that the method of recording is different, e.g. the explanation of the exercise could have been done differently by different physicians.

3. Each doctor follows the predefined exercise protocol differently.

4. Overlying characteristics of the exercises matches between certain PG's. This can also be seen in the number of double exercises found in a single file in 7

5. Not making use of PG 4 is resulting in less training in the overlying characteristics.

6. Only using the 5 exercises that all patients have in common, compared to the previous group, who used all available data,

results in less training in the overlying characteristics or maybe they influenced the model by cleaning the data by hand.

Conclusion

Given the research question: “To what extend and in what way, can different supervised data science techniques be used on kinematic recordings to contribute to a more valid and more reliable diagnosis, made by a doctor, on shoulder disability.”

The answer is: ”Using normalization and a low frame count per exercise gave the most accurate and precise results. Combining data verification and cleaning to verify that no assumptions are made with training the model. It shows classification can be done on kinematic recordings.”

According to the results, the cleaning of the dataset influenced multiclass classification with a logistic regression model on kinematic recordings in a positive way. While making use of data cleaning methods e.g.; normalizing, splitting double exercises in two files and only using 5 frames per exercise, improves results for a simple logistic regression model (69% accuracy). Other data cleaning measures that were taken, e.g. remove idle, caused lowered results (65% accuracy). This gives the hint, that the Logistic Regression model might be too simple for the complexity of the kinematic recordings, since removing unnecessary data gets rid of bias in the data set. An example could be that the different PG’s didn’t have the same physician doing the recording, this could lead to longer or shorter idle times before and after the exercise in each PG. This point is also proven by the fact, that the last research on this subject, didn’t use any data cleaning methods that were used in this research, which gave a way higher accuracy (97%) (Kasper van der Hoofd, 2019) over all 4 PG’s.

Possible flaws in this research are, that there is no certainty on why, PG 2 matches with PG 3 but not vice versa. It also could be, that the labels for the PG’s are not correct, therefore giving wrong results. That being said, the results of this research show that it is safe to say that it is possible to do multiclass classification on kinematic recordings by making use of a machine learning model by making use of Logistic Regression. The possibility to make use of a more complex model such as a Convolutional Neural Network, in order to get even better classification in PG 2 and 3 in the future is seen.

References

Bevel, p. (2020, 01 15). *thefouriertransform.com*. Retrieved from thefouriertransform.com: http://www.thefouriertransform.com/

C.G.M. Meskers, H. F. (1999). Calibration of the “Flock of Birds” electromagnetic tracking device and its application in shoulder motion studies. *ELSEVIER*, 629-633.

C.G.M.Meskers, H. V. (1998). 3D shoulder position measurements using a six-degree-of-freedom electromagnetic tracking device. *ELSEVIER*, 280-292.

D. Douglas Miller, E. W. (2018). Artificial Intelligence in Medical Practice: The. *The American Journal of Medicine*, 129/133.

Hoffman, J. I. (2019). Basic Biostatistics for Medical and Biomedical Practitioners (Second Edition). *Academic Press*, 581-589.

Kasper van der Hoofd, L. d. (2019). *Kinematic analysis of shoulder motion for diagnostic purposes.* The Hague.

LUMC . (2020, 15 01). *Leids Universitair Medisch Centrum*. Retrieved from Laboratory of Kinematics and Neuromechanics (LK&N): https://www.lumc.nl/org/orthopedie/OnderzoekenInnovatie/81211001022579/LKN/

LUMC. (2020, 01 16). *Leids Universair Medisch Centrum*. Retrieved from Dr. ir. JH (Jurriaan) de Groot: https://www.lumc.nl/org/revalidatiegeneeskunde/medewerkers/111231015815519

Maaten, L. v. (2014). Accelerating t-SNE using Tree-Based Algorithms. *Journal of Machine Learning Research 15*, 3221-3245.

Sylvia A. Stegeman, P. B. (2016). Journal of Electromyography and Kinesiology. *ELSEVIER*, 74-80.