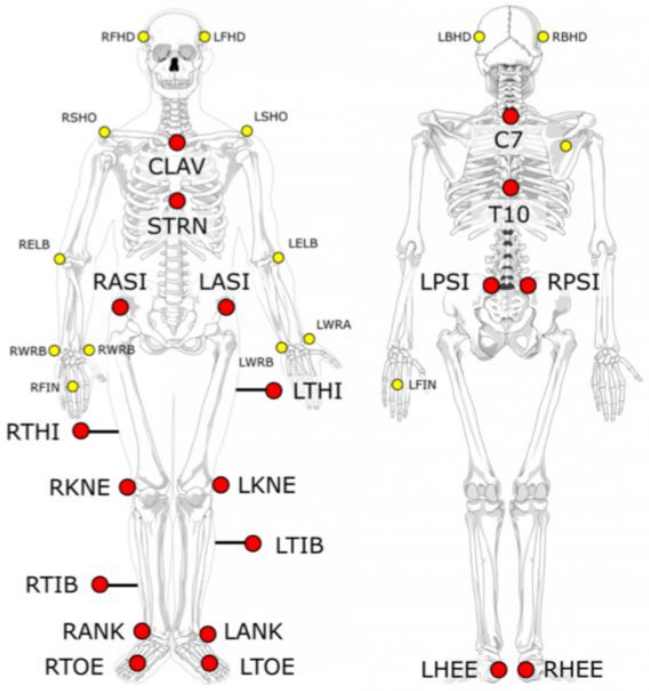
Predicting Events for Gait

Data Mining - Final Project Report

Kieran Schubert and Marc Desaules



June 14th 2019

Summary

An investigation to find a suitable approach and to build a model that predicts as correctly as possible the events along the gait cycle of a random patient has been carried out. The first approach, which tries to directly predict the events with the correct frame has the caveat to not have enough training data. For this reason, work has been done on the given data with the software “Mokka” to complete the dataset with more events. This additional work has made it possible to improve the misclassification rate and the number of events detected. However, to try and reach better predictions and based on this first analysis, a second approach was taken, still with the aim of increasing the amount of data on which the model can train. The data have been relabeled to have a model that predicts the stance and the swing of the patient’s gait instead of the “foot strike” and the “foot off” events. This second approach with a Neural Network prediction method gives, for the random test set, a score of 2 “fake” events and an average frame error prediction of 1.6.

Introduction

Patients suffering from a pathological gait have the opportunity to improve their condition by going through a surgical operation. For the surgeon in charge of the operation, the type and the degree of the pathology of a patient are important information used to adjust the surgical intervention. The actual process for obtaining more information on the patient pathology is to carry out a physical examination and a clinical gait analysis. The physical examination measures different characteristics such as the range of motion, the “stiffness” of the muscles, or the dimensions of the body. The clinical gait analysis is the subject of interest in this project and we are asked to develop an algorithm that will help to obtain the patient gait cycle by detecting automatically the different phases of the gait with the information gathered by cameras and markers on the patient’s body.

Problem Description

The clinical gait analysis consists of making the patient walk on a 10 meter long walkway and recording the physical motion behaviour with the help of the markers and the cameras. Figure 1 below shows the position of the markers on the patient’s body.

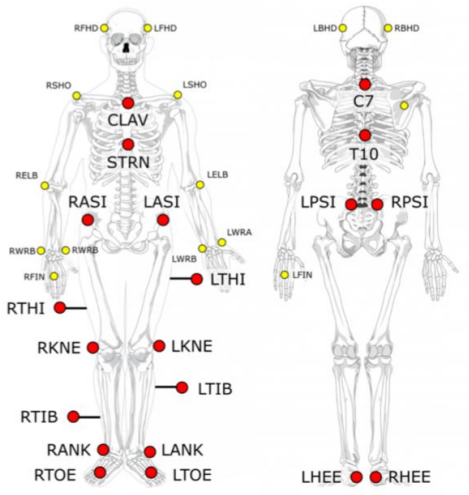


Figure 1 : Position of the markers on the patient body

In addition, a platform with force sensors detects different events of the patient’s gait. The four events detected are the strikes by the right and the left foot and when each foot takes off the ground. These events constitute the four phases of the human gait cycle as shown in Figure 2 below:

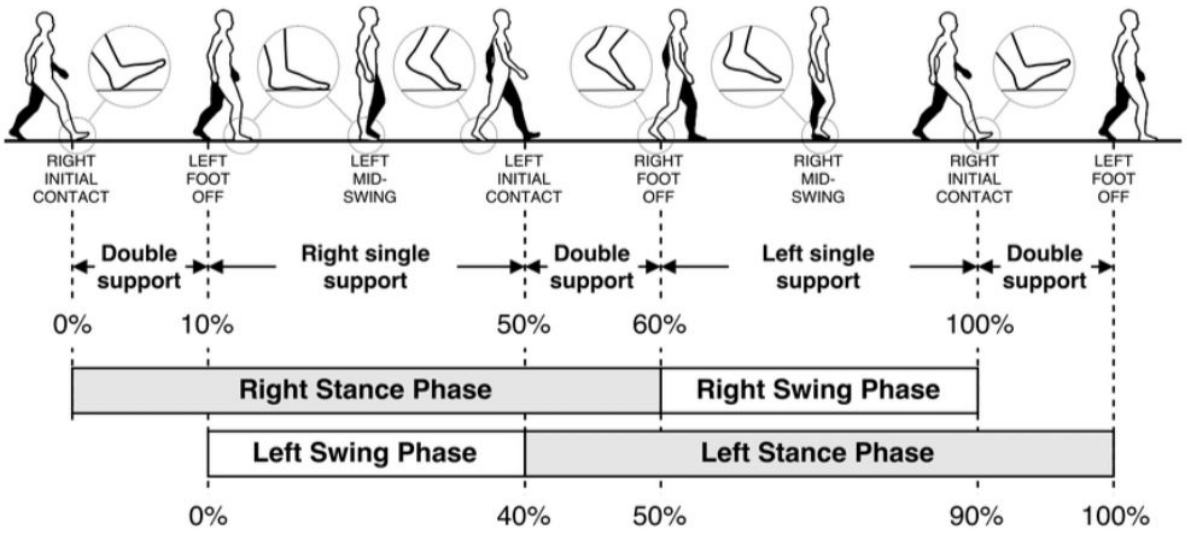


Figure 2: The human gait cycle

The problem is that the detection of the four events is made by a signal analysis and depends on the quality of the signal. Often, not all the events are detected and the completion has to be done manually or the patient needs to walk on the walkway again. To avoid this waste of time and make the process easier for the patient and the medical personnel, it would be more convenient to get the events directly from the markers through a statistical model.

Strategy

There are several ways to approach this problem and we have chosen two approaches for which we have tested different statistical methods that we will discuss in the next sections.

The first approach was to consider the gait events to predict directly at the frames at which the events happen. This means that all the frames are considered equal to a “null” event, except the frames with an event (i.e. Foot Strike Right, Foot Strike Left, Foot Off Right, Foot Off Left). The advantage of this approach is that we have directly the frame with the predicted event and we can take advantage of the percentage human gait framework as shown in figure 2 to improve the model. On the other hand, there is limited information on which the machine can learn to predict the events since there is only few frames with an event per patient (sparsity and unbalanced labels).

The second approach was to consider, in a first step, each frame with a value to predict and in a second step, identify the events by detecting the change between the stance and swing phases. The advantage of this approach is that we can use all the frames to train the model. However, it

is necessary to recode the output of the dataset with the values for each frame (i.e Right Stance Phase, Right Swing Phase, Left Stance Phase, Left Swing Phase).

In both approaches, the more information we have on the events, the better the result, and since the data provided contains all the markers, it was possible to complete the missing events by hand with the provided software Mokka. Before modelling the data with different methods, we had to go through a data processing to collect, understand and format the dataset.

Data Processing

Data collection

The data was provided by the Sofamehack Challenge as '.c3d' files. A total number of 90 patients with a maximum number of 35 markers on their bodies were recorded walking on a 10 meter segment. As each patient possesses the minimal markerset (20) but some can have a few more (35), we had to decide which markers to extract. In the first approach, we used the maximum number of common markers for each patient (33) as predictors. In the second approach, markers related to the head have been omitted and we uses only 29 markers. Finally, these data were extracted using the BTK module in Python 2.7 and saved in a '.csv' format after the data formatting process.

Data structure

All patients were recorded walking for a certain number of frames and gait events were identified using a force plate. However, not all events were detected, hence the dataset was initially incomplete with regards to the labels. The predictors consist of 33 body markers, each marker containing 3 spatial coordinates.

	First approach	Second approach
Number of patients	90	90
Type of the data	c3d format	c3d format
Number of predictors	99 (numerical)	87 (numerical)
Labels	5 levels (categorical)	4 levels (categorical)
Patient pathologies	Cerebral Palsy (CP), Foot Disorder (FD), Idiopathic Toe Walker (ITW)	Cerebral Palsy (CP), Foot Disorder (FD), Idiopathic Toe Walker (ITW)

Table 1: Structure of the data

Data formatting

The data of all patients was merged into a single '.csv' file of size (23'452, 101), with one column of the dataset containing the 5 response labels coded as follows:

- 0: No gait event
- 1: Foot Strike Left
- 2: Foot Strike Right
- 3: Foot Off Left
- 4: Foot Off Right

The first column of the dataset contains the 'Frame number', and the next 99 columns contain the 33 body markers for each (x, y, z) spatial dimension. The last column contains the labels.

An important issue is that the labels are very unbalanced (see figure 3). The majority of labels are '0', whereas the counts of labels we are interested in identifying are very low (only 3.9% of labels are non-zero).

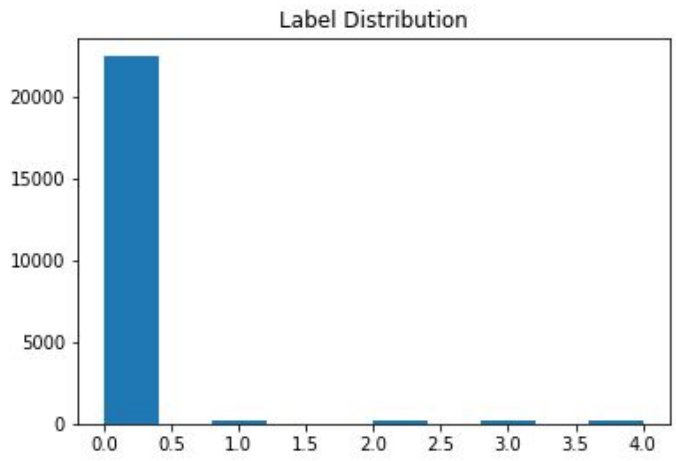


Figure 3: Label Distribution

0	Foot Strike Left	Foot Strike Right	Foot Off Left	Foot Off Right
22'538	234	230	225	225

Table 2: Label Counts

Different strategies were implemented to resolve this issue.

Exploratory Data Analysis

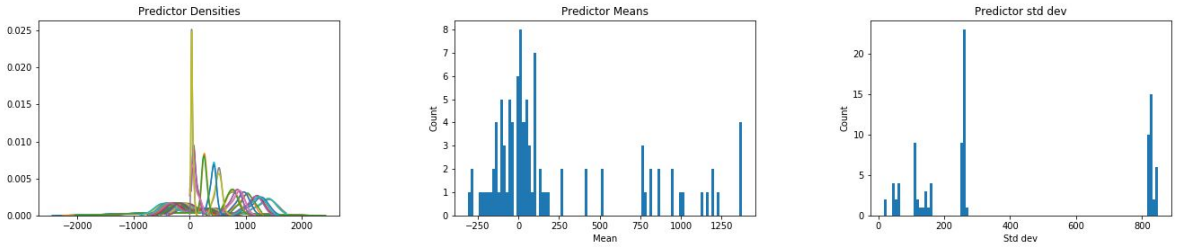


Figure 4: Predictor distribution

A quick look at the data shows that the predictors do not seem to be iid, but do seem to be Gaussian. The means and standard deviations are quite different, thus some modelling will require feature standardization.

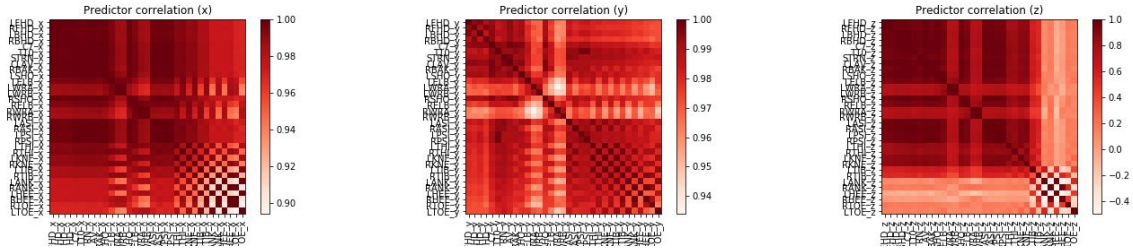


Figure 5: Predictor correlation

Predictors appear to be highly correlated between themselves in the same dimension, but not between dimensions (not shown here). Thus, we assumed there existed no multicollinearity between the predictors.

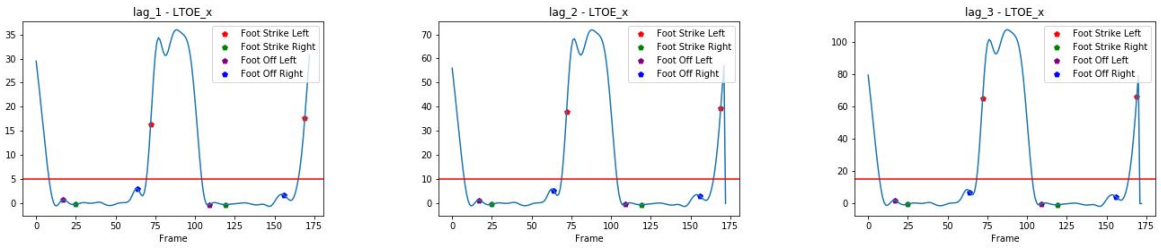


Figure 6: Predictor behaviour around gait events

Some markers exhibit particular behaviour around gait events, and this feature will be exploited later on.

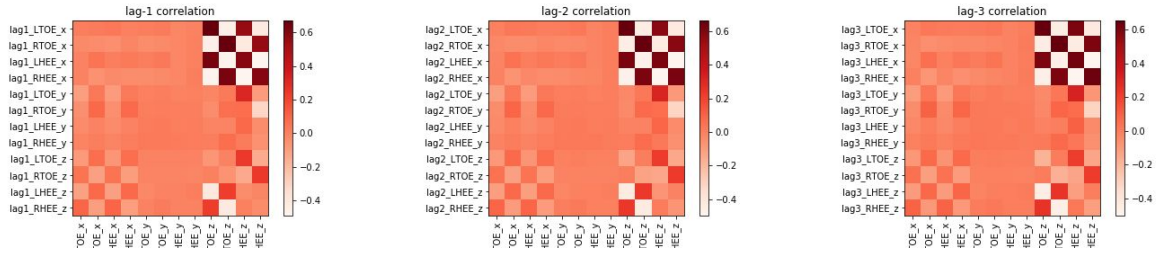


Figure 7: Lag and predictor correlation

Different time lags were computed on specific predictors. The previous plots show no important linear association between the lags and predictors, except for the lags associated to their respective predictors. Thus the dataset was augmented with the lags without fear of multicollinearity.

First approach

Strategy description

The first approach consisted of 2 modelling phases based on two different datasets. The first phase of modelling was conducted on the entire raw dataset (hereby *dataset 1*) with the following 99 predictors (each marker with 3 spatial coordinates) :

LFHD	RFHD	LBHD	RBHD	C7	T10
STRN	CLAV	RBAK	LSHO	LELB	LWRA
LWRB	RSHO	RELB	RWRA	RWRB	LASI
RASI	LPSI	RPSI	LTHI	RTHI	LKNE
RKNE	LTIB	RTIB	LANK	RANK	LHEE
RHEE	RTOE	LTOE			

Table 3: 33 markers with three dimensions values each

This dataset is of dimension (23’452, 101), with the labels and frame numbers.

Then, lags of 1, 2 and 3 for each (x, y, z) dimension were computed on the following predictors to include the time-series aspect of the data, adding 36 predictors to the dataset:

LTOE	RTOE	LHEE	RHEE

Table 4: Markers used for lags

This phase of feature engineering seemed justified, as the manual coding of gait events using Mokka showed some particular behaviour of these predictors in gait event identification (see figure 6). Indeed, one can observe that the lagged data for the ‘LTOE_x' at lags 1, 2, 3 exhibits a particular cyclic behaviour. In the case of a ‘Foot Off Left’ event, we see that the value of ‘LTOE-x’ at all 3 lags quickly attain a minimum. Furthermore, the events ‘Foot Off Left’, ‘Foot Off Right’ and ‘Foot Strike Right’ all occur under a certain threshold. This information can help during modelling or at a later stage to refine predictions. Since the labels were very sparse (not many labels different from 0), a “padding” approach was undertaken. For each patient, the 3 frames before and after a non-0 event were padded with the same event to try and balance the distribution of labels. This dataset is of dimension (23’452, 137), and is referred to as *dataset 2*.

The Results section shows that although the results based on an accuracy metric are in some cases quite satisfactory, the fact that the labels are heavily unbalanced pushes the models to mostly predict the majority class of '0' and no events of interest.

Consequently, the second modelling phase was conducted using a sampling technique to balance the classes in the training and test sets, and analysis was done with the predictors of the augmented *dataset 2*.

Statistical methods

The following methods were used on the unbalanced *dataset 1* and *dataset 2*:

- Gaussian Naive Bayes, kNN, Softmax Classification, SVM, LDA, QDA, Neural Network (3 models)

The following method was used on the balanced *dataset 2*:

- Neural Network

All models were trained on a Training Set (80%) and evaluated on a Test Set (20%), as well as with 5-fold Cross Validation (when computationally feasible) on the entire Training Set, using classification accuracy as an evaluation metric.

Model Specifications:

Model	# Classes	Distribution				
GNB	5	Gaussian				

Model	NN	Distance				
kNN	5	Euclidean				

Model	Max it.	Solver				
Softmax	1e6	lbfgs				

Model	Kernel	C	Gamma	Shrinkage		
SVM	Gaussian	1	$n^{-1}\Sigma$	Yes		

Model	Hidden layers	Nodes per layer	Optimizer	Mini-Batch	Epochs	Loss	Activations
NN 1	1	8	SGD	256	100	X-entropy	ReLu, Softmax

Model	Hidden layers	Nodes per layer	Optimizer	Mini-Batch	Epochs	Loss	Activations
NN 2	2	10	SGD	256	100	X-entropy	ReLu, Softmax

Model	Hidden layers	Nodes per layer	Optimizer	Mini-Batch	Epochs	Loss	Activations
NN 3	4	100	SGD	256	100	X-entropy	ReLu, Softmax

Table 5: Model specifications

Results

Unbalanced datasets

Model	**Test accuracy [%] Dataset 1**	**Correct non '0' predicted [%] Dataset 1**	**5-fold CV accuracy (+/- std) Dataset 1**	**Test Accuracy [%] Dataset 2**	**Correct non '0' predicted [%] Dataset 2**	**5-fold CV accuracy (+/- std) Dataset 2**
GNB	60	2.2	65 (+/- 11)	56	24	55 (+/- 18)
kNN	96	0	96 (+/- 0)	77	0	65 (+/- 7)
Softmax	96	0	NA	82	54	NA
SVM	96	0	NA	82	62	NA
LDA	96	9.3	85 (+/- 7)	79	43	75 (+/- 12)
QDA	69	1.5	89 (+/- 5)	77	45	73 (+/- 8)
NN 1	96	0	96 (+/- 0.2)	83	52	83 (+/- 0.9)
NN 2	96	0	96 (+/- 0.1)	84	54	84 (+/- 0.5)
NN 3	96	0	96 (+/- 0.2)	90	74	91 (+/- 0.6)

Table 6: Model results for the unbalanced datasets

Balanced dataset

Model	**Test Accuracy [%] Dataset 2**	**Correct non '0' predicted [%] Dataset 2**	**5-fold CV accuracy (+/- std) Dataset 2**
NN 3	99	99	98 (+/- 0.2)

Table 7: Model results for the balanced *dataset 2*

The results from the first modelling phase are very clear. Although high accuracy results are obtained on the raw *dataset 1*, many of the predicted events are '0'. We do not obtain relevant predictions throughout the models. By augmenting the dataset with lags as well as padding the labels around non '0' events, we increase performance significantly. Indeed, even though accuracy decreases for all models, the percentage of correctly predicted non '0' events increases dramatically. A 4 hidden layer Neural Network (*NN 3*) with 100 nodes per layer achieved the best results with 91% CV accuracy, and 74% non '0' events detected.

In the second modelling phase, the augmented *dataset 2* was resampled in order to create balanced classes. In this scenario, the training set was comprised of around 14'500 training examples for each class, totaling 72'540 samples in the Training Set. The best performing Neural Network from the previous modelling phase was used and achieved 98% CV accuracy and predicted 99% of non '0' events, which were the best results obtained.

As a result, we obtained a very performant model for predicting specific gait events, although the event padding procedure reduces the frame accuracy of predictions. In further steps, it would be possible to use the cyclic behaviour of certain markers to refine predictions around events in a post-modelling phase.

Second approach

Strategy description

To take advantage of all the frames information, this second approach uses the fact that the human gait cycle has four phases. The data collected through the markers should be strongly related to the position of the patient's body in space. Therefore, it should have a strong relationship between the markers and the different stance and swing phases. The hypothesis is that if it is possible to predict the phase of gait in which the patient is at each frame, then we are able to predict the events knowing each phase switch. To do so, we must first apply a new label on each frame in our data set as shown in figure 3 below.

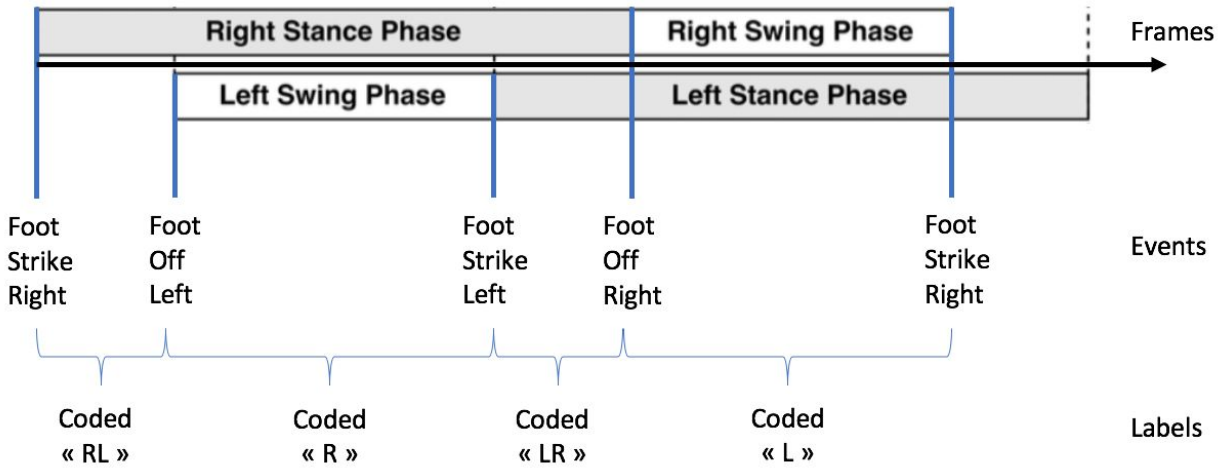


Figure 8: New labels for the frames regarding the events

The markers are the predictors and for this second approach we used 29 markers per patient.

C7	T10	STRN	CLAV	RBAK	LSHO
LELB	LWRA	LWRB	RSHO	RELB	RWRA
RWRB	LASI	RASI	LPSI	RPSI	LTHI
RTHI	LKNE	RKNE	LTIB	RTIB	LANK
RANK	LHEE	RHEE	RTOE	LTOE	

Table 8: 29 markers with three dimensions values each

Each marker has three values for the three dimensions axes which makes in total 87 predictors.
After the data processing, we ended with a dataset which contains the frame number, the predictors, the original labels, the new labels and contextual data. The table 2 below shows a summary of the dataset.

Frame	C7x1	T10x1	.....	RTOEx3	LTOEx3	Label	Context_x	Event	Group	Patho	Patient	Year	Month	Day	Trial	Classes
20	-1029	-1018		44.96	49.898			0	CP	GMF CS1	1972	2015	2	2	19	R
21	-1023	-1009		45.141	48.425	Foot Strike	Left	1	CP	GMF CS1	1972	2015	2	2	19	LR
22	-1017	-1001		45.435	47.293			0	CP	GMF CS1	1972	2015	2	2	19	LR
.....																

Table 9: Final dataset before applying statistical methods

The contextual data, such as the patient number and the patient trial, are important in order to subsequently analyse the predictions' efficiency and confront them with the markers through the software "Mokka".

Statistical methods

Before carrying out prediction with a statistical model, we split the dataset in a training (80%) and a test set (20%). This split is done at the level of the patients and not at the frame level. In order to test different statistical methods on the training set and be able to choose the best one, we also have created a cross-validation function with 5 folds that computes misclassification error mean and the standard deviation of the mean. The figure 4 below shows the different split percentage.

Figure 9: Percentage of the dataset splits

In addition, it is possible that the original dataset contains missing values. When the missing value is in the training set, the frame is deleted and when the missing value is in the test set, it is replaced by the mean of the same predictor values. Finally, the predictors are standardized to avoid scale biases.

We tried different statistical methods to model the data as the following:

- Naive Bayes - kNN - Random Forest - Gradient Boosting - Neural Network

Results

Misclassification rates

In the second approach, the misclassification rate measures the percentage of frames that are misclassified. This rate is related to the new classes of labels and not to the events. Nevertheless, as we will see, there is a strong relationship between this performance indicator and the score that measures the accuracy of predicting an event correctly. The misclassification rate is computed for the cross-validation function, which is an average on the 5 folds for the best model (i.e best hyper-parameters chosen) and it is computed for the test set where the best model has been trained on the complete training data set.

Scores

The predictions have to meet two requirements. The first one is to detect an event at the correct frame and the second one is to identify the correct event (i.e Foot Strike Right, Foot Off Right, Foot Strike Left, Foot Off Left). This is the reason why in Table 3 below, we have included a score that counts the number of events detected with the correct label and an associated score that counts the number of frames on average between the true values and the predictions on the events detected correctly. To have an idea of the number of "fake" events detected, we add the number of frames predicted in total.

Table of results

Test set: 4183 frame;168 events; 17 patients

Methods	Hyper-parameters	CV Average Misclass. % error	Test set Misclass. % error	Test set Score Number of events detected [-10,10]	Test set Score Average Frame error for the detected events
Naive Bayes	-	36.8%	33.78%	104/168 (161)	3.84
kNN	k=50	30.8%	20.67%	102/168 (139)	3.5
Random Forest	Max_features = 30	20.2%	11.59%	162/168 (186)	2.49
Gradient Boosting	learning _rate=0.1 max_depth=4	18.8%	15.6%	149/168 (176)	3.06
Neural Network	Alpha = 1 Hidden layers size=(500,200,10) solver='sgd	13.2%	6.38%	169/168 (170)	1.59

Table 10: Results for the different models

Our best model is the Neural Network method which predicts 2 “fake” events and when the events are correctly labeled, the prediction error on the frames is on average around 1.6 frames. However, the test set gives a good misclassification rate compared to the misclassification rate of the cross-validation function and we should be aware that on average the results is more around 13% than 6%, which corresponds to a test score of 3.23. In addition, to compute the score, the 10 first and the 10 last frames of each patient have been omitted.

There are still several opportunities to improve the prediction score of the Neural Network model. Further investigations can be carried out to select better hyper-parameters such as the number of layers or the number of nodes. In addition, to avoid “fake” events, it is possible to constraint the predictions to follow the specific pattern of the human gait cycle. Finally, techniques such as variable selection could help to decrease the variance of the predictions.

Conclusion

The investigation to find a suitable approach and to build a model that predicts as correctly as possible the events along the gait cycle of a random patient have been carried out. The first approach, which aims at predicting directly the events with the correct frame has the caveat of not having balanced labels and correct labelling. For this reason, work has been done on the given data with the software “Mokka” to complete the dataset with more events. Using an augmented dataset with lags and event padding, it was possible to reach high prediction accuracies with Neural Networks. Then, by resampling the augmented dataset in order to balance the labels, the best results of this first approach were achieved. However, further refinements and post-modelling could be undertaken in the future to fine-tune the final model.

However, to try to reach better predictions and based on this first analysis, a second approach was taken, based on the same idea of increasing the number of data on which the model can train. The data have been relabeled to have a model that predicts the stance and the swing of the patient’s gait instead of the foot strike and the foot off event. This second approach with a Neural Network prediction method gives, for the random test set, a score of 2 “fake” events and an average frame error prediction of 1.6. Further investigations to improve the average score are possible by working on the model parameters, the variable selection or by constraining the prediction to a specific pattern.