A Machine Learning Approach to Automated Gait Analysis for the Noldus Catwalk System

Holger Fröhlich, Kasper Claes, Catherine De Wolf, Xavier Van Damme, and Anne Michel

***Abstract*—*Objective*: Gait analysis of animal disease models can provide valuable insights into in vivo compound effects and thus help in preclinical drug development. The purpose of this paper is to establish a computational gait analysis approach for the Noldus Catwalk system, in which footprints are automatically captured and stored. *Methods*: We present a - to our knowledge - first machine learning based approach for the Catwalk system, which comprises a step decomposition, definition and extraction of meaningful features, multivariate step sequence alignment, feature selection, and training of different classifiers (gradient boosting machine, random forest, and elastic net). *Results*: Using animal-wise leave-one-out cross validation we demonstrate that with our method we can reliable separate movement patterns of a putative Parkinson's disease animal model and several control groups. Furthermore, we show that we can predict the time point after and the type of different brain lesions and can even forecast the brain region, where the intervention was applied. We provide an in-depth analysis of the features involved into our classifiers via statistical techniques for model interpretation. *Conclusion*: A machine learning method for automated analysis of data from the Noldus Catwalk system was established. *Significance:* Our works shows the ability of machine learning to discriminate pharmacologically relevant animal groups based on their walking behavior in a multivariate manner. Further interesting aspects of the approach include the ability to learn from past experiments, improve with more data arriving and to make predictions for single animals in future studies.**

***Index Terms*—Animal behavior, gait recognition, machine learning.**

I. Introduction

ANIMAL models, most often rodents, play an important role in drug development. They are used to test compounds before applying them in clinical trials in order to evaluate potential efficacy for a specific disease and minimize risks for humans. In that context also analysis of animal movement and behavioral patterns in response to a specific treatment is of high interest. For example, Fröhlich *et al.* [1] used Support Vector Machine classifiers to discriminate rats treated with antidepressant drugs from controls based on video recordings of their swimming behavior. Other authors have focused on gait analysis systems involving the detailed study of animal locomotion, mostly via video recordings and classical univariate statistics [2], [3]. Following this line Vandeputte *et al.* [4] demonstrated a first gait analysis of a putative rat model for Parkinson's (PD) and Huntington's (HD) disease. Recently, also machine learning techniques have been applied for gait analysis, mostly in human [5]–[7].

The focus of this paper is on a computational gait analysis of a putative mouse PD model. PD is a progressive neurodegenerative movement disorder. The clinical symptoms are bradykinesia, rigidity and resting tremor, which are caused by the degeneration of dopaminergic neurons in the substantia nigra pars compacta, leading to the subsequent depletion of dopamine levels in the striatum [8]. Accumulation of misfolded alpha-synuclein (α-syn) into Lewy Bodies and neurites is one of the pathophysiological hallmark of PD. Self–amplification, propagation and transmission of misfolded α-syn are hypothesized as major contributing factors to disease onset and progression [9]. Site specific injection of the neurotoxin, 6-hydroxydopamine (6-OHDA), or the fibrillar form of α-syn (synthetic Pre Formed Fibrils: PFFs) within the dopaminergic structures of the brain are commonly used to induce degeneration and/or α-syn pathology in wild-type mice [10], [11]. In mice, unilateral injection of 6-OHDA into the striatum produces severe dopamine depletion and measurable forelimb akinesia [12]. By contrast, unilateral injection of PFFs in the same structure induces α-syn pathology but a rather unilateral gradual loss of the dopaminergic neurons [11]. It is therefore controversially discussed whether motor and behavioral abnormalities could be observed.

The goal of this study was to evaluate whether motor deficits in mild parkinsonian animals could be computationally detected and discriminated from those observed with severely impaired mice. To this end we developed a - to our knowledge first - machine learning based approach for the Noldus Catwalk gait analysis system[1]. Catwalk is a gait analysis system for mice and rats, in which animals traverse a glass plate voluntarily towards a goal box. During the run footprints are automatically captured and stored. Within our machine learning based approach we first defined and extracted a large set of potentially relevant features from these recorded raw data. Afterwards we used minimum-redundancy-maximum-relevance based feature selection [13] in combination with several classifiers (elastic net [14], Random Forests [15], Gradient Boosting Machine [16]) to learn the discrimination of movement patterns of animal with different brain lesions (Sham, PFF) in two brain regions (substantia nigra and striatum). We demonstrate the prediction performance of our approach via animal-wise leave-one-out cross-validation in

Manuscript received March 22, 2017; accepted April 25, 2017. Date of publication August 24, 2017; date of current version April 19, 2018. *(Corresponding author: Holger Fröhlich.)*

H. Fröhlich is with the UCB Biosciences GmbH, Monheim 40789, Germany (e-mail: holger.froehlich@ucb.com).

K. Claes, C. D. Wolf, X. V. Damme, and A. Michel are with the UCB Pharma SA.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

Digital Object Identifier 10.1109/TBME.2017.2701204

[1] http://www.noldus.com/animal-behavior-research/products/catwalk

dependency of time elapsed since the initial treatment. Moreover, we show that our system is able to differentiate highly accurate between movement patterns of animals with the same leason type at different time points after intervention. Finally, we demonstrate that with our modeling approach we can discriminate 6-OHDA treated animals and several control groups with high prediction accuracy. We analyzed the change of this performance in dependency of the number of animals included into the study. A further contribution of our work is an in-depth analysis of the features involved into all our classifiers via statistical techniques, which helped model interpretation.

Altogether our approach demonstrated the usefulness of multivariate machine learning techniques to analyze the complex data retrieved from the Catwalk gait analysis system, which in turn may help to better understand movement patterns in animal experiments.

II. Material and Methods

A. Animal Experiments

All animal experiments were performed according to the Helsinki declaration and conducted in accordance with the guidelines of the European Community Council directive 2010/63/EU and Belgian legislation. The ethical committee for animal experimentation from UCB Biopharma SPRL (LA1220040 and LA 2220363) approved the experimental protocols. C57BL/6J mice (Charles River France) were housed in cages for one week before experimentation. They were kept on a 12:12 light/dark cycle with light on at 06.00 AM and at a temperature maintained at 20–21 °C, at humidity of approximately 40%. All animals had free access to standard pellet food and water before assignment to experiments. Additional enrichment and welfare were provided (Enviro-dri PharmaServ) before and after the surgery. Animal health was monitored daily by the animal care staff. Surgeries were performed under ketamine and xylazine anesthesia, and all efforts were made to minimize suffering. Sacrifice was done with CO2. A total of 14 μg of 6-OHDA (3.5 μl, 0.5 μl/min) was injected at two different sites into the right striatum at the following coordinates: AP: $+1.0$, $+0.3$; ML-2.1, -2.3; DV: -3.2, -3.2. A total of 5 μg of sonicated α-syn PFFs (2 μL, 0.1 μL/min) was injected into the right striatum or the right substantia nigra at the following coordinates: AP: 0.2; ML: -2.0; DV: -3.0 (striatum) and AP: -3.0; ML: -1.3; DV: -4.35 (substantia nigra). Sham mice were brain injected with vehicle at the same coordinates.

Two series of experiments with the Noldus Catwalk system were conducted. First, a batch of female mice were tested at 23 weeks post-surgery and were assessed according to the following four conditions: vehicle (Sham), 6-OHDA, murine PFFs and non-lesioned. A second batch of male mice were followed over 6-month period and assessed in the Catwalk system at 7, 11 and 25. In this second set of experiments, four different lesion conditions were assessed: vehicle in the striatum (Sham striatum), murine PFFs in the striatum (PFF striatum), vehicle in the substantia nigra (Sham substantia nigra) and murine PFFs in the substantia nigra (PFF substantia nigra). The idea behind not mixing the two genders in the same experiment was (1) to avoid some potential variability associated with gender difference and, (2) to avoid any behavioral contamination (e.g. by increased arousal).

Each test in the Catwalk system was repeated at least three times. Each Catwalk test was done in two phases. First, an habituation period wherein the mice were allowed to walk around the testing environment for five minutes. As a second step, the animals were tested in the system during a 4-day period, at maximum, in order to get three consecutive optimal crossings of the corridor. At each testing session, mice were habituated to the testing room for 30 minutes and the system was cleaned between each animal tested.

Table I gives an overview about all performed experiments and corresponding animal groups.

TABLE I
Overview About Animal Experiments and Groups

Experiments	Groups
7 weeks, male	PFF striatum ($n = 17$), Sham striatum ($n = 16$)
11 weeks, male	PFF striatum ($n = 16$), PFF subst. nigr. ($n = 17$), Sham striatum ($n = 17$), Sham subst. nigr. ($n = 16$)
25 weeks, male	PFF striatum ($n = 17$), PFF subst. nigr. ($n = 16$), Sham striatum ($n = 16$), Sham subst. nigr. ($n = 16$)
23 weeks, female	non-lesioned ($n = 10$), PFF ($n = 9$), 6-OHDA ($n = 17$), Sham ($n = 9$)

Numbers in brackets indicate the number of animals in each group. Each Catwalk run test was repeated at least three times, i.e. there were more data samples than animals.

B. Catwalk Gait Analysis

Raw footprint features recorded for each animal run by the Noldus Catwalk system at a sampling rate of 100 Hz comprise:

1) print length
2) print width
3) print area
4) mininum intensity
5) maximum intensity
6) mean intensity

These are features for each of the four paws, which are recorded over the length of the whole animal run (Fig. 1). Obviously, if a particular paw does not touch the ground, then the value of these features for that paw is not defined/missing.

C. Feature Definition and Extraction

1) Global Features: A number of global characteristics were extracted from the raw Catwalk data explained before. These are summarized in Table II and already gave rise to more than 500 features. The majority of these features captured auto-correlations of raw footprint features for different paws (6 raw footprint features × 21 different time lags × 4 auto-correlations for 4 paws). Calculation of these auto-correlations omitted missing values. The animal weight was only used to correct for possible biases in a later step (see below).

2) Step Decomposition and Extracted Step Features: Signals for each paw were next decomposed into individual steps. A step in that context was defined as the time period during which a particular paw was on ground. For each step 15 different groups of features (147 in total) were extracted (Table III), comprising general characteristics of the step (such as step length and duration) as well as descriptors of the signal distribution and frequency spectrum (one for each raw footprint feature). One of step feature was also the estimated true mean of the signal, corrected for the influence of other paws on ground. The rational was that e.g. footprint intensity of a particular paw depends on which other paws are on ground at the same time.

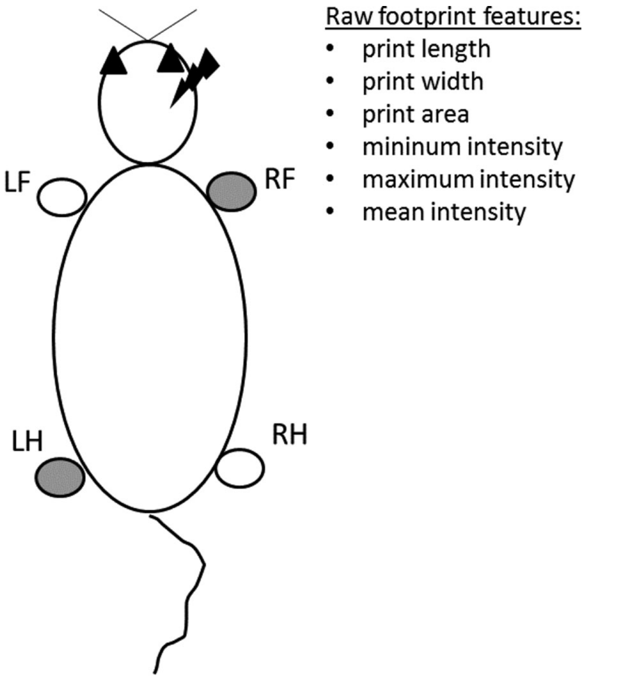

Fig. 1. Schematic view on raw footprint features captured by the Noldus Catwalk system in combination with the nomenclature for paws that we use in this article. In the shown example the gray filled RH and LF indicate that both paws are on ground at the same time and corresponding raw footprint features are captured. Brain lesions were always applied to the right hemisphere of animals in our data.

TABLE II
OVERVIEW ABOUT GLOBAL FEATURES EXTRACTED FROM THE CATWALK RAW DATA

Global feature group	Total	Description
run direction	1	start of animal run from left or right in Catwalk system
paw usage frequencies	4	percentage of time, in which each of the 4 paws is on floor
paw transition frequencies	16	percentage of time, in which after paw X there is a measurement for paw Y (or paw X again)
opposite paws on ground	2	percentage of time, in which opposite paws (LF + RH, RF + LH) are on floor
auto-correlations	504	Auto-correlations (time lags −0.1, −0.09, ..., 0.1 s) calculated separately for each raw footprint feature and paw

TABLE III
OVERVIEW ABOUT FEATURES EXTRACTED FOR EACH STEP

Step feature group	Total	Description
step length	1	step length
step duration	1	time duration of step
step speed	1	ratio of step length and duration
mean	6	signal mean (one for each raw footprint feature)
variance	6	signal variance (one for each raw footprint feature)
estimated true mean	6	estimated signal mean, corrected for the influence of other paws on ground (one feature for each raw footprint feature)
0%, 25%, 50%, 75%, 100% quantiles	24	quantiles of signal distribution (one for each raw footprint feature)
skewness	6	skewness of signal distribution (one for each raw footprint feature)
kurtosis	6	kurtosis of signal distribution (one for each raw footprint feature)
entropy	6	Shannon entropy of signal density (one for each raw footprint feature) using kernel density estimation (R function ’density’)
relative paw contribution	6	mean relative contribution of paw to total signal, summed over all paws (one for each raw footprint feature)
auto-correlation	30	signal auto-correlation (time lag = 0.01, ..., 0.05 s); one variable for each raw footprint feature
spectral entropy [17]	6	Shannon entropy of spectral density
adaptive sine multitaper power spectral density [18]	36	total power within frequency bands [0, 0.1 Hz], (0.1, 0.2 Hz], ..., (0.4, 0.5 Hz]; one variable for each frequency band and raw footprint feature
amplitude	6	maximum amplitude of frequency spectrum (one variable for each raw footprint feature)

Most of the listed features are calculated for each raw footprint feature.

Parts of the signal variability within each step can be explained by movement of other paws. In order to correct the measured signal y_{fp} for paw p and footprint feature f for these confounding effects we decomposed y_{fp} according to a standard ANOVA model:

$$y_{fp} = \beta_0 + \sum_{q \neq p} \beta_q x_q + \epsilon_{fp}$$

where x_q represents a 0/1 indicator for paw q being on ground at the same time as p, β_0, β_q are regression coefficients and ϵ_{fp} the measurement noise. The model can be fitted via ordinary least squares to the available measurements from each step. The intercept β_0 represents the desired estimate for the expected true signal, since it reflects the fraction of the signal variance that cannot be attributed to any of the other paws $q \neq p$.

In order to achieve a comparison of different animal runs (comprising different number of steps) we aggregated step features separately by their median over the first and last three steps as well as of the rest (i.e. the middle) of the run. This was done in order to account for the possibly different behavior of the animal at the beginning and end of a run experiment. In addition we also considered the overall distribution of each step feature, described via the 0%, 25%, 50%, 75% and 100% quantile.

3) Step Sequence Alignment via Multivariate Dynamic Time Warping: In addition to the above mentioned aggregation of individual step features over the whole length of a run we applied an alignment of steps from individual paws. This was done via Dynamic Time Warping (DTW) in a multivariate manner [19]. More specifically, for each pair of paws we first computed an Euclidean distance matrix of steps based on the step features defined in the last Section. Then an open ended, asymmetric alignment of steps was applied, see [19] for details. The result was an *asymmetric* DTW dissimilarity between each pair of paws. That means the alignment of e.g. the right hind

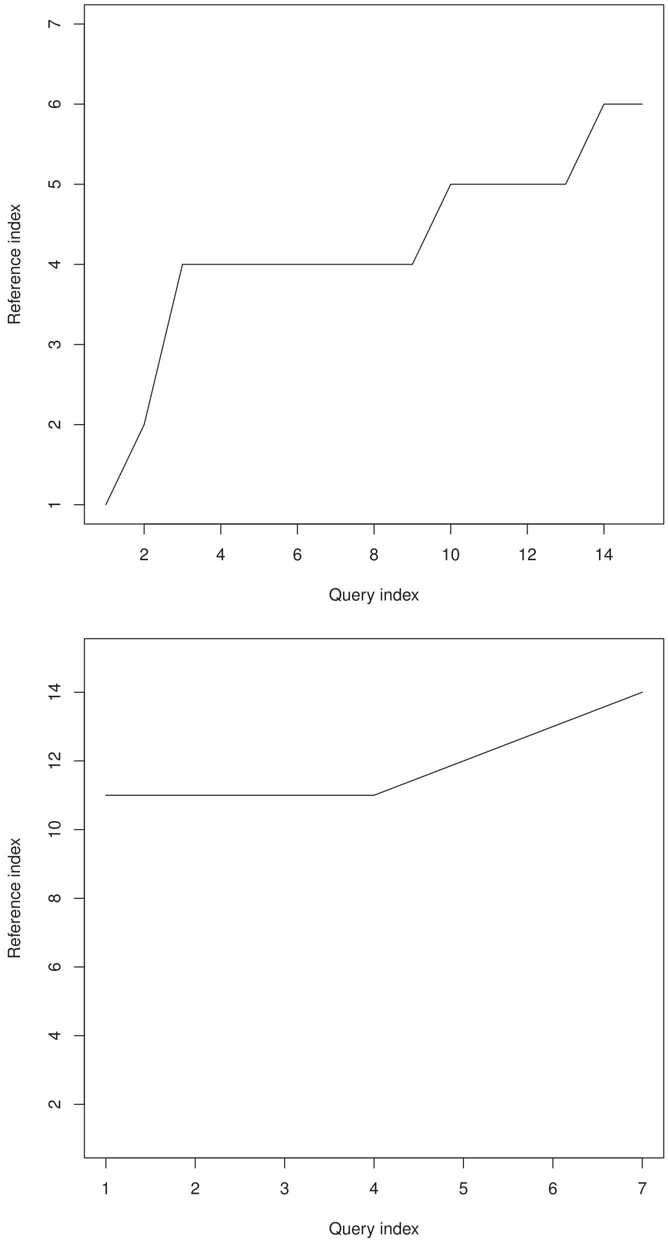

Fig. 2. Example of open ended, asymmetric DTW based alignment of two step sequences. Top: alignment of first versus second sequence. Bottom: vice versa. The figures depict the step indices which are mapped to each other.

step sequence to the left hind step sequence is not necessarily the same as vice versa (Fig. 2). We decided for this type of alignment, because step sequences of two runs could be of different length and only match in certain parts of the steps sequences.

In addition to aligning steps sequences between paws from the same animal run we also aligned step sequences from the same paw between different animal runs. The result was a DTW dissimilarity between each pair of animal runs in the dataset. For a given candidate animal we then calculated the average DTW dissimilarity to the nearest $k = 3, 5, 7$ animal runs from each animal group. This was done while leaving out other samples from the same candidate animal in order to avoid overoptimistic results. Notably, we expect step sequences from the same animal over different runs to be more similar than step sequences from different animals. Different values of k were chosen to capture more local (smaller value) and more global dissimilarities (larger value) while at the same time accounting for the typical animal group sizes in our experiments.

TABLE IV
Overview About Features Extracted From Step Sequence Alignments

Step Sequence Dissimilarity	Total	Description
DTW dissimilarities within animal run	12	alignment of step sequences for each pair of paws X, Y within same animal run: one alignment for X against Y and one for Y against X
DTW dissimilarities across animal runs	$12G$	alignment of step sequences of the same paw across different animal runs: average DTW dissimilarity to $k = 3, 5, 7$ nearest animal runs from each group

In conclusion of the step sequence alignment $12G + 12$ features were extracted, were G denotes the number of animal groups (Table IV).

D. Correction for Weight Bias

Animals typically gain weight over time, and we cannot exclude the possibility that different brain lesions may influence that weight gain differently (see Figs. 1–4 in Supplements Material). This would imply a bias in our downstream analysis, because our goal is to discriminate different animal groups based on their movement patterns and not based on their weight. Hence, we removed all features showing a squared Spearman rank correlation larger than 0.25 with weight (including weight itself). Likewise, all features containing missing values were omitted.

E. Machine Learning

Based on the above defined features a machine learning classifier was used to predict the class of an animal in a multivariate fashion. We applied Random Forests (RF) [15], elastic net (EN) [14] and a Gradient Boosting Machine (GBM) [16] for that purpose. EN was chosen as a popular example of a linear classifier that is sparse (i.e. keeps only few non-zero coefficients) and tends to select groups of correlated features together. It should be mentioned that our data is high dimensional, i.e. contains far more variables than samples. Moreover, between these variables non-trivial correlations are expected. EN is a machine learning method that has been developed to exactly address this situation. RF on the other hand are a very popular decision tree based ensemble learning technique that enables non-linear classification of data with heterogeneous variables [20]. That means RF appropriately deal with features having different numeric scales and distributions - a situation that also appears in our application. However, with a high fraction of irrelevant variables (which in a high dimensional setting is expected) the classification performance of RF is typically degrading. Therefore, we chose as a third technique a GBM, which is a decision tree based ensemble learning method that is more costly to train than RF, but in many applications found more robust and predictive [21].

While RF rely on the idea of iteratively re-sampling the data with replacement and growing decision trees with large depth (bagging), the GBM constitutes a weighted ensemble of weak decision tree classifiers with restricted maximal depth (here 5). Therefore, a GBM typically results into sparse model with fewer selected than available features - a property that is desirable for high dimensional data.

We applied EN, RF and GBM as follows: While iterating through all animals in the data we sequentially left out all samples (i.e. runs) from one particular animal, trained a classifier based on the remaining samples and then made predictions for the left out animal runs. Note there were data from at least three runs of the same animal. Overall classification performance was assessed via the accuracy (i.e. fraction of correctly classified samples) and by complete confusion tables.

To address class imbalance we used sample and class weights. The weight for each sample was defined as one divided by the relative size of the respective animal class of the sample. Similarly class weights for RF were defined. Regularization parameters for EN and the optimal number of boosting steps for GBM were found via an inner cross-validation loop during the validation procedure outlined above. The overall number of features in our data far exceeds the number of samples. Hence, for computational speed up and to reduced the overfitting risk we per-filtered features according to the minimum redundancy maximum relevance method [13]. This was done within the leave-one-animal-out validation procedure outlined above. Only the top 500 features were kept for classifier training. Notably, EN and GBM are not necessarily using all of these features.

F. Classifier In-Depth Analysis

In addition to prediction performance, from an application perspective an important aspect is the interpretation of machine learning classifiers. After having assessed the prediction performances of different classifiers as described in the last Section the best classifier was trained on the whole dataset and variable importance estimated. For GBM this was done by calculating the relative loss reduction [16]. For RF the mean decrease in prediction accuracy was calculated via a permutation test [15]. For EN the fraction of non-zero coefficients was investigated.

To aid further interpretation we then grouped primary features into several overlapping classes (Table S6) and asked, whether features within a specific class were present more often than expected by chance in the final classifier. To answer this question we performed a hyper-geometric test, which assesses the statistical over-representation of selected variables within a specified class relative to a background. In our case the background consisted of all variables in the dataset. The output of the hyper-geometric test for each feature group is a p-value, which we corrected for multiple testing with dependency using the false discovery rate (FDR) control method by Benjamini and Yekutieli [22].

In addition we investigated, the tendency of features in a certain group to discriminate a specific pair of animal classes. For that purpose we first conducted a Wilcoxon test for each individual feature in the final classifier and ranked variables according to the absolute value of the test statistic. We then asked whether features from a certain group were predominantly appearing at the top of the ranked list. This was answered with the help of a Gene Set Enrichment Analysis (GSEA), which is a well known technique for interpretation of biological high throughput data [23]. Briefly, GSEA successively calculates a running sum statistic over a ranked list, which is increased if a feature is within a certain group and decreased otherwise. The magnitude of the increment depends on the correlation of the feature with the animal class. The running sum test statistic corresponds to a modified Kolmogorov-Smirnov statistic, and GSEA estimates the statistical significance via a permutation test (here: 10,000 permutations), resulting into a p-value. A significant p-value corresponds to a clear deviation from the null hypothesis that variables in the feature group of interest have no stronger tendency to separate animal groups than any other random feature group of the same size. We corrected the p-value for multiple testing (because there are multiple variable groups) in the same way as described above.

It should be noted that for both analysis tasks (statistical over-representation and tendency for pairwise discrimination) a non-significant p-value does not necessarily imply feature redundancy or irrelevance. A significant over-representation of features from a certain groups points towards an enrichment of discriminatory information in that group. Likewise, a significant GSEA result indicates a tendency of features in a certain class to separate two animal groups individually, i.e. without the help of other features. The opposite reasoning is not true. That means a non-significant result does not indicate a lack of discriminatory information in a certain feature group.

TABLE V
ACCURACIES FOR PREDICTING THE CORRECT LESION TYPE AT DIFFERENT TIME POINTS

Groups	#animals	#runs	EN (%)	RF (%)	GBM (%)
PFF (striatum), Sham (striatum) at week 7	33	99	**96%**	**96%**	**96%**
PFF (SN), Sham (SN), PFF (striatum), Sham (striatum) at week 11	66	237	37.1%	67.5%	**78.9%**
PFF (SN), Sham (SN), PFF (striatum), Sham (striatum) at week 25	65	195	32.8%	48.2%	**51.3%**

III. Results

A. Different Lesion Types are Predictable From Movement Patterns

In a first set of experiments we asked, how well different machine learning classifiers (RF, EN, GBM) could discriminate between PFF injected and Sham animals, where the lesion was applied in the substantia nigra (SN) and the striatum, respectively. Table V summarizes overall accuracies for animals 7, 11 and 25 weeks after lesion. GBMs showed the overall highest prediction performance with 96% accuracy at week 7, 79% at week 11 and 51% at week 25. Confusion tables for best classifier models are shown in Tables S1–S3.

The GBM classifiers at weeks 7 and 11 allowed for a clear separation of interventions in different brain regions. Specifically, PFF interventions in striatum and SN were clearly discriminated. At week 25 only Sham interventions in the striatum and PFF in the SN were rather clearly separable from the others, whereas the other two interventions (Sham in the SN and PFF in the striatum) were mostly indistinguishable from the

TABLE VI
ACCURACIES FOR PREDICTING THE CORRECT TIME POINT AFTER BRAIN LESION

Groups	#animals	#runs	EN (%)	RF (%)	GBM (%)
Sham (striatum) at weeks 7, 11, 25	34	102	**98%**	97.3%	96.7%
PFF (striatum) at weeks 7, 11, 25	45	135	**98.4%**	96.8%	**98.4%**
Sham (SN) at weeks 11 and 25	32	96	**99%**	**99%**	97.9%
PFF (SN) at weeks 11 and 25	33	99	**100%**	**100%**	**100%**

TABLE VII
CONFUSION TABLE SHOWING THE PREDICTION PERFORMANCE OF A GBM MODEL DISCRIMINATING 6-OHDA MEDICATED AND CONTROL ANIMALS

	truth			
predicted	non-lesioned	PFF	6-OHDA	Sham
non-lesioned	63.3%	0%	3.9%	0%
PFF	0%	88.9%	0%	0%
6-OHDA	33.3%	11.1%	96.1%	7.4%
Sham	3.3%	0%	0%	92.6%

The overall accuracy was ~87%

others. Altogether it seems that observable movement patterns are dependent on the lesion type and location. Furthermore, there might be a recovery from or adaptation to lesions over time, which dependends on the exact lesion type (Sham or PFF) and location (striatum or SN). This conditional development of animals might explain the drop in prediction performance at week 25.

When analyzing the final GBM model trained on the whole dataset at week 7 in more detail we found a clear over-representation of right fore, footprint shape and middle of run related features (Table S7).

At week 11 74% of the cumulative importance was contributed by features resulting from step sequence alignment from different animals (Table S8). These features were at the same time over-represented (FDR $<$ 1E-4). Moreover, there was an over-representation of left hind related features (9% cumulative importance) and features related to the middle of animal runs (5% cumulative importance). Individual animal groups were in tendency separated by right fore (Sham vs. PFF in striatum, PFF in SN vs. PFF in striatum,) as well as footprint intensity related features (Sham vs. PFF in SN).

At week 25 88% and 65% of the cumulative importance was contributed by step features and the statistical distribution of them over the length of an animal run (Table S9). These features were at the same time over-represented (FDR $<$ 1E-4 and $<$1E-6, respectively). Furthermore, there was again an over-representation of left fore related features (24% cumulative importance) and features related to the middle of animal runs (19% cumulative importance). Individual animal groups were in tendency separated by right hind and fore (Sham vs. PFF in striatum, PFF in SN vs. PFF in striatum), left fore (Sham vs. PFF in SN, Sham in SN vs. Sham in striatum), footprint intensity (Sham vs. PFF in SN) related features as well as feature auto-correlations (Sham vs. PFF in striatum, PFF in SN vs. PFF in striatum) and step speed (PFF in SN vs. PFF in striatum).

B. Movement Patterns Change Over Time

We asked whether animals with the same lesion type could be separated with respect to the time passed since the intervention (i.e. 7, 11 and 25 weeks) based on their movement behavior. We conducted the same computational validation procedure outlined above. The results indicate very clear differences in the treatment effects over time, partially without any classification error (Table VI, Tables S4–S6). The best prediction performances where overall achieved with EN.

In all cases a high fraction of selected variables were step features (more than 90%; Tables S11–S14). Moreover, footprint shape and intensity related features were contained in all EN models. Except for the small EN model discriminating PFF in SN at weeks 11 and 25 (having only 9 non-zero coefficients) there was always an over-representation of features related to the middle of animals runs. In the EN model discriminating Sham in SN at weeks 11 and 25 there was in addition an over-representation of right hind and left hind (33% and 34% of all features) as well step length related features. For animals with Sham intervention in the striatum left fore, footprint intensity and footprint shape had a tendency for discriminating weeks 25 vs. 11, whereas for weeks 11 vs. 7 footprint shape and left hind related features were significant. For animals with PFF injection in the striatum weeks 25 vs. 11 were in tendency separated by step features, footprint shape, left fore and auto-correlation related features. For weeks 11 vs. 7 we found a significance of right fore related features.

C. 6-OHDA and Control Animals can be Well Discriminated

Finally, we asked, whether our method could discriminate PFF injected animals from non-lesioned, Sham treated, and 6-OHDA treated 23 weeks old male animals. The same protocol to estimate prediction performance as before was used. We obtained an accuracy of 86.7% using GBM, followed by 77.8% and 54.1% with RF and EN, respectively. The confusion table (Table VII) specifically showed a very clear separation of 6-OHDA, Sham and PFF animals. The prediction performance for non-lesioned animals was a bit lower than for the other groups.

Analysis of the final GBM model revealed a high cumulative influence and statistical over-representation of features related to the step sequence alignment from different animals (55% cumulative influence, Table S15). Further notable aspects were the relatively high cumulative influence and over-representation of right hind (47% cumulative influence) and footprint intensity related features (19% cumulative influence). Furthermore, there was an over-representation of features related to the middle of an animal run (FDR = 0.03).

D. Learning Curve

To obtain an idea, how much the results shown in the last Section depend on the number of animals included into the data we left out all samples from 1, 3, 5 randomly picked animals

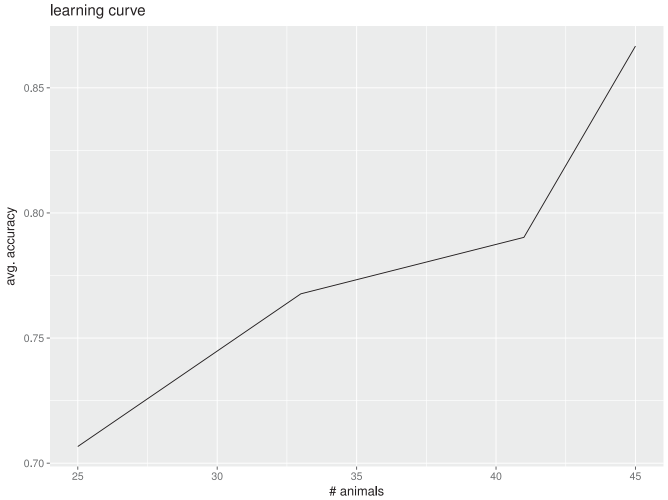

Fig. 3. Learning curve showing prediction performance as a function of the number of animals in the dataset. Prediction performance at each data point was always assessed with a leave-one-animal-out procedure.

from each group. For each of these sub-sampled datasets we then re-ran the complete analysis to estimate the prediction performance five times. The learning curve generated in this way (Fig. 3) showed a decrease from ∼87% to ∼70%, if the number of animals dropped from 45 to 25.

IV. Conclusion

The present work showed the ability of a machine learning system to discriminate pharmacologically relevant animal groups based on their walking behavior. On the technical side one of the key innovations of our work is the definition and extraction of relevant features from raw Catwalk data. As it stands, our approach thus primarily relies on this specific gait system. However, our feature extraction and selection approach as well as machine learning methods could in principle also be adapted to other gait systems.

We suggested an approach for in-depth analysis of the complex classifiers used in our system. All of the employed classifiers combined and integrated a large number of different features to reach their prediction performance. This is in contrast to classical univariate statistical analysis methods [2], [3]. Moreover, a key difference to these methods is that our system is able to make predictions. Hence, the trained model could be used in future experiments to characterize the probability of an animal to fall into one of the existing groups. This would be doable even for one single animal. That means the information contained in past animal experiments can be re-used, which implies a gain in efficiency and is also ethically beneficial. Notably, the existing classifier models can be updated with every newly conducted animal experiments. Hence, prediction performance is expected to increase beyond the level shown here. In Section III-D we have tried to quantify this behavior via a learning curve. Altogether we think that our system provides a step towards a better understanding and automated gait analysis in animal experiments.

References

[1] H. Fröhlich *et al.*, “Automated classification of the behavior of rats in the forced swimming test with support vector machines,” *Neural Netw.*, vol. 21, no. 1, pp. 92–101, 2007.

[2] K. A. Clarke and J. Still, “Gait analysis in the mouse,” *Physiol. Behavior*, vol. 66, no. 5, pp. 723–729, Jul. 1999.

[3] P. Yu *et al.*, “Gait analysis in rats with peripheral nerve injury,” *Muscle Nerve*, vol. 24, no. 2, pp. 231–239, Feb. 2001.

[4] C. Vandeputte *et al.*, “Automated quantitative gait analysis in animal models of movement disorders,” *BMC Neurosci.*, vol. 11, p. 92, Aug. 2010.

[5] Q. Riaz *et al.*, “One small step for a Man: Estimation of gender, age and height from recordings of one step by a single inertial sensor,” *Sensors*, vol. 15, no. 12, pp. 31 999–32 019, Dec. 2015.

[6] R. Joyseeree *et al.*, “Applying machine learning to gait analysis data for disease identification,” *Studies Health Technol. Informat.*, vol. 210, pp. 850–854, 2015.

[7] T. Shirakawa *et al.*, “Gait analysis and machine learning classification on healthy subjects in normal walking,” *AIP Conf. Proc.*, vol. 1648., Mar. 2015, Art. no. 580009.

[8] J. M. Beitz, “Parkinson’s disease: A review,” *Frontiers Biosci. (Scholar Ed.)*, vol. 6, pp. 65–74, Jan. 2014.

[9] H. Braak *et al.*, “Stages in the development of parkinson’s disease-related pathology,” *Cell Tissue Res.*, vol. 318, no. 1, pp. 121–134, Oct. 2004.

[10] R. K. Schwarting and J. P. Huston, “The unilateral 6-hydroxydopamine lesion model in behavioral brain research. analysis of functional deficits, recovery and treatments,” *Progress Neurobiol.*, vol. 50, no. 2/3, pp. 275–331, Oct. 1996.

[11] K. C. Luk *et al.*, “Pathological α-synuclein transmission initiates parkinson-like neurodegeneration in nontransgenic mice,” *Science*, New York, NY. USA, vol. 338, no. 6109, pp. 949–953, Nov. 2012.

[12] M. Lundblad *et al.*, “A model of l-dopa-induced dyskinesia in 6-hydroxydopamine lesioned mice: Relation to motor and cellular parameters of nigrostriatal function,” *Neurobiol. Disease*, vol. 16, no. 1, pp. 110–123, Jun. 2004.

[13] L. Yu and H. Liu, “Efficient feature selection via analysis of relvance and redundancy,” *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

[14] J. Friedman *et al.*, “Regularization paths for generalized linear models via coordinate descent,” *J. Statist. Softw.*, vol. 33, no. 1, pp. 1–22, 2010, [Online]. Available: http://www.jstatsoft.org/v33/i01/

[15] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[16] J. H. Friedman, “Stochastic gradient boosting,” *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002.

[17] J. Jung and J. D. Gibson, “The interpretation of spectral entropy based upon rate distortion functions,” in *Proc. 2006 IEEE Int. Symp. Inf. Theory*, Jul. 2006, pp. 277–281.

[18] A. J. Barbour and R. L. Parker, “psd: Adaptive, sine multitaper power spectral density estimation for R,” *Comput. Geosci.*, vol. 63, pp. 1–8, Feb. 2014.

[19] Toni Giorgino, “Computing and visualizing dynamic time warping alignments in R: The dtw package,” *J. Statist. Softw.*, vol. 31, no. 7, pp. 1–24, 2009.

[20] T. Hastie *et al.*, *The Elements of Statistical Learning*. New York, NY, USA: Springer, 2001.

[21] R. Caruana and A. Niculescu-Mizil, “An empirical comparison of supervised learning algorithms using different performance metrics,” in *Proc. 23 rd Int. Conf. Mach. Learn*, 2005, pp. 161–168.

[22] Y. Benjamini and D. Yekutieli, “The control of the false discovery rate in multiple testing under dependency,” *Ann. Statist.*, vol. 29, pp. 1165–1188, 2001.

[23] A. Subramanian *et al.*, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.” *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 43, pp. 15 545–15 550, Oct. 2005.