Identification and Interpretation of Gait Analysis Features and Foot Condition by Explainable AI

Mustafa Erkam Özates
Fraunhofer IPA
Alper Yaman (alper.yaman@ipa.fraunhofer.de)
Fraunhofer IPA
Firooz Salami
Heidelberg University Hospital
Sarah Campos
Heidelberg University Hospital
Sebastian I. Wolf
Heidelberg University Hospital
Urs Schneider
Fraunhofer IPA

Research Article

Keywords: Gait Analysis, Machine Learning, Explainable AI, Classification, Decision Support Systems, Feature Extraction, Foot Conditions

Posted Date: October 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2187167/v1

Abstract

Background

Clinical gait analysis is a crucial step for identifying foot disorders and surgery planning. However, a large amount of gait data makes this assessment difficult and time-consuming. There are separate efforts to reduce its complexity by manually or automatically finding features (e.g. minimum of a joint angle in a specific axis), identifying the foot condition by Machine Learning (ML), and interpreting the outcome by explainable artificial intelligence (xAI).

Methods

In this article, we explore the potential of state-of-the-art ML algorithms to automate all these steps for a set of 6 foot conditions. New features are created manually and then recursive feature elimination is employed based on Support Vector Machines (SVM) and Random Forest (RF) to eliminate the features with low variance. SVM, RF, K-nearest Neighbor (KNN), Logistic Regression (LREGR), and Majority Voting (MV) algorithms are compared for classification and Local Interpretable Model-agnostic Explanation (LIME) is used for the interpretation of the outcome of the ML models. 40 features are eliminated and 334 features are given to the classifier models as inputs.

Results

The foot conditions are classified with a maximum average accuracy of 0.86 by KNN and MV, maximum average recall of 0.97 by KNN, and max average F1 score of 0.86 by KNN and MV.

Conclusions

High success scores indicate that the relation between the selected features and foot conditions should be strong and meaningful, potentially indicating clinical relevance. All models are interpreted for each foot condition for random 20 patients and the most contributing features are graphically demonstrated. The proposed ML pipeline can be easily extended for other foot conditions and retrained as new data arrives. It can help experts and physicians in the identification of foot conditions and the planning of potential surgeries.

1. Background

Gait analysis provides a large amount of high-dimensional 3D data (e.g. joint angles and ground reaction forces) after post-processing the images from well-calibrated cameras and sensors (1). The gait data is obtained by averaging among strides, as a result, giving the mean and standard deviation values along the whole gait cycle. It is hard to evaluate since it should be investigated manually by experts and

physicians. Despite the data being quantitative, its evaluation is mostly based on the experience and expertise of the experts and the patient demography they are familiar with. Consequently, their evaluation contains human bias inherited in its nature. Despite computer-based evaluation tools help to handle large data, calculating major features (e.g. minimum or a maximum of a specific joint angle) can reduce the data dimensions and make it easier to explain. Automatic feature generation could be implemented using mathematical tools, e.g. principal component analysis (PCA) (2) or linear discriminant analysis (LDA). Note that these features may not correspond to kinematic characteristics of the joints and hence may not have clinical relevance.

It is possible to identify foot conditions by either assessing the whole gait data (i.e. time series normalized to a gait cycle) or the generated features. Besides, identification of foot conditions can be automated by utilizing state-of-the-art Machine Learning (ML) based classification methods. This automation may help experts to speed up the assessment of the gait data.

Artificial Intelligence (AI), and specifically ML has dramatically become popular within the last decades since it shows its maturity in prediction in several areas, especially when the tasks allow automation. Furthermore, ML methods have been widely used in medical diagnosis and treatment to automate time-consuming tasks as well as make a better and faster diagnosis and treatment plan. In many medical fields, AI proved its potential in diagnosis, even better and faster than human experts (3, 4).

Alongside the success of ML methods in classification, it is difficult to interpret their outcome because of their black-box nature (5). This limits the application of ML approaches in medical fields since the reasons for predictions should be known for further diagnosis and treatment. Explainable Artificial Intelligence (xAI) has drawn more attention due to the need to interpret ML models. Shapley Additive Explanations (SHAP) (6) and Local Interpretable Model-agnostic Explanation (LIME) (7) are two of the common xAI methods that help to interpret the predictions.

We believe that the ML approach is capable of automating feature selection and foot condition identification to support experts and physicians in their assessment. For this purpose, in this study, the state-of-the-art feature elimination and ML algorithms were explored and facilitated the diagnosis of various foot conditions. Furthermore, an xAI method was exploited to interpret the outcomes of the ML models. An ML pipeline was built that automates feature selection, foot condition identification, and interpretation by xAI.

Related Work

Wolf et al. presented an automated feature assessment workflow specifically for gait analysis of cerebral palsy (CP) patients (8). This approach consists of deriving time series from the original time series, computing scalar features from the derived and original time series, and evaluating computed scalar features (9). ML methods were successfully employed in clinical gait analysis (10, 11) for the classification of cerebral palsy (12), osteoarthritis (13), multiple sclerosis (14), etc. Layer-Wise Relevance Propagation (LRP) was used to interpret the prediction of the individual gait pattern classification model

(15). Slijepcevic et. al. used LRP to explain the classification results of lower-body gait disorders (16). To the best of our knowledge, this study is the first to combine feature selection, ML classification, and xAI to use gait data and identify foot conditions. Doederlein et al. developed graphical decision matrices of foot deformities to assist physicians in disease discrimination for their diagnostic decision processes (17–21).

2. Methods

2.1. Subjects

The anonymized retrospective gait data of 248 patients with 6 different foot conditions and 100 subjects with the typically developed feet has been collected in the course of patient care over the last two decades in the Department of Orthopedics and Traumatology, Heidelberg University, Heidelberg, Germany to be used in the project mentioned in the Acknowledgments Section. The Heidelberg Foot Measurement Method (HFMM) (8) was employed for data collection using a 12-camera VICON motion capturing system (Vicon Motion Systems Ltd. Oxfordshire UK) (Table 1).

Table 1
Foot conditions, number of subjects, and age

Foot conditions (Classes)	Number of subjects	Age (mean ± std)	Height(cm) (mean ± std)	Weight(kg) (mean ± std)
Tibiotalar osteoarthritis + partial ankle replacement	58	57.9 ± 11.7	170.5 ± 8.6	82.6 ± 17.7
Planovalgus	64	31.3 ± 15.6	171.9 ± 12.4	71.7 ± 20.1
Consolidated calcaneal fracture	20	52.5 ± 10.4	178.0 ± 8.2	86.6 ± 11.7
Hallux rigidus	40	58.6 ± 8.2	168.2 ± 9.6	74.8 ± 13.3
Club foot	41	11.8 ± 9.3	140.1 ± 28.2	41.5 ± 21.8
Cavovarus	25	19.5 ± 14.9	161.0 ± 18.9	59.7 ± 23.1
Typical feet	100	24.0 ± 15.2	159.8 ± 23.7	55.6 ± 23.9

2.2. Dataset and Functional Angles

The number of subjects among foot conditions varies significantly (Table 1), which indicates that the dataset is highly imbalanced. For each subject, the dataset contains the processed data of 12 functional angles; that was averaged across 7–10 strides for each subject and normalized to a percentage gait cycle. Therefore, for each functional angle, mean and standard deviation (std) time series were calculated, each consisting of 101 data points representing the gait cycle from 0% to %100. The non-

averaged data could be used directly that results in 7–10 times more data. It will be evaluated in future studies after the first findings are gathered within the proposed study.

The functional angles, described by Simon et. al. (8) are as follows: Tibio-talar flexion, Medial arch inclination, Medial arch angle, Lateral arch angle, Subtalar inversion, Forefoot/ankle supination, Forefoot/midfoot supination, Forefoot/hindfoot abduction, Forefoot/ankle abduction, Inter MT I-V angle, Hallux adduction, Hallux flexion.

2.3. Feature Extraction and Selection

The feature extraction and selection steps are denoted in Fig. 1. The time series were used similar to the automated feature assessment workflow of Wolf et al. for deriving new time series (i.e. the first gradient and difference from normative) and computing scalar features of both original (i.e. mean and std time series) and derived ones (9). Subsequently, median imputation and normalization were performed to fill the missing values and rescale the values between 0 and 1. A feature elimination method was applied before ML classifications since some of the features might have low variance and therefore, negligible effect in classification. To do so, ML-based feature selection algorithms were applied. Below, the feature selection and elimination steps are explained in detail.

2.3.1. New Time Series Derivation

Two new time series were derived from the mean time series for each functional angle and each subject: The first gradient time series and the difference from the normative time series.

The first gradient" V " of the mean time series "U" was calculated according to the Formula (1):

$$V[k] = \frac{1}{2}(U[k+1] - U[k-1])$$

1

A discrete derivation of the mean time series must be done since the time series are in a discrete domain. "k" is the data point index within the time series.

The difference relative to a reference considered to be normal U_{norm} namely difference from normative " DN " was calculated according to the Formula (2):

$$DN[k] = |U[k] - U_{norm}[k]|$$

2

For each functional angle, reference normal time series U_{norm} was calculated by averaging the corresponding time series across all typical subjects.

2.3.2. Feature Computation

The original and derived time series either could be the inputs of the learning algorithm. However, computing scalar features from these time series significantly decreases the input size causing lesser but more representative data to be learned. This dimension reduction reduces the complexity of both the input data and the model.

Segmenting the time series regarding gait phases (stance and swing) ensures extracting more specific information from the time series since a comparison of peak values and other computed features for each gait phase should be physiologically more meaningful (as suggested by Wolf et al. (9). For this purpose, the mentioned features were computed both for the stance and swing phases of the gait cycle, except the range of motion feature, which was calculated for the whole gait cycle.

For each functional angle, the computed scalar features from both derived and original time series were as follows: Minimum and maximum values and their timings (i.e. temporal position in the gait cycle; x-axis in Fig. 2). Note that, for the std time series, only the maximum value and its timing were considered as features since it makes sense that some disorders may cause more differences in the gait patterns among the strides. For the first gradient time series, the average difference from normative was additionally considered because the difference from the normal gait pattern may contain meaningful information. As seen in Table 2, each functional angle had 15 features for stance and swing phases separately. The number of features reach 31 when the features were calculated for stance and swing phases separately (15x2 = 30), and the range of motion was added as another feature. The number of extracted features for all functional angles can be calculated as (number of functional angles)x(number of features) = 12x31 = 372. Additionally, the “Foot_off” scalar value from the dataset was added which denotes the time when the foot is off to divide the stance and swing phases. There were 2 “Foot_off” values; namely mean and standard deviation among the strides for each leg. Thus, the number of the whole features for each subject reached 374.

Table 2
Computed features for each time series

Time series	Max value	Timing max	Min value	Timing min	Average difference from normative
Mean	X	X	X	X	
Std	X	X			
First gradient	X	X	X	X	
Difference from normative	X	X	X	X	X

2.3.3. Filling the Missing Values and Normalization

A median imputation for filling the missing values of a subject was implemented using the other subjects of the same disorder, similar to the mentioned one in a study about missing longitudinal data (22).

The input values were of different scales and had to be normalized for simplifying the learning process of the AI algorithms. Normalizing input values between 0 to 1 scale did ease the learning process of the algorithms (23). Moreover, a natural logarithmic transformation was applied to skew the input distribution to the normal distribution. However, no significant improvement through skewing was observed.

2.3.4. Feature Selection

Computed scalar features for all original and derived time series ended up with 374 scalar values per subject. Not just for computational applicability, but also for problem-related specification (24), the most useful subset of features was selected. Before applying two feature elimination methods, only the data of the affected legs were selected from the dataset.

Recursive feature elimination with cross-validation algorithm of Scikit-Learn module (25) was used for feature selection. This algorithm was implemented with a Support Vector Machine (SVM) (26) algorithm as an estimator, which was trained recurrently with the newest subset of the features, the least important of which (regarding model coefficients) was then eliminated for the next trial. The feature subset securing the maximum accuracy in the estimator algorithm was then assumed to be the ideal subset according to this algorithm. In addition to this algorithm, a Random Forest (RF) (27) model was trained with all computed features and the importance parameter (25) within the trained model was used to select the most relevant features by eliminating the ones below a threshold. The intersection of the selected relevant feature subsets of these two algorithms was used as the optimum subset for training the classifier models. For this optimum subset, 334 of the 374 features were selected, which was the intersection of the outcomes of the aforementioned algorithms.

2.4. Classification for the Identification of Foot Conditions

Figure 1 shows the classification steps. Briefly, the dataset was split for training and testing and median imputation and normalization were applied to each subset, separately. In training, hyper-parameters of the ML models were optimized and then models were trained. In testing, the trained models were tested with the split unseen data.

2.4.1. Dataset Splitting

The dataset was split into training and testing groups with a ratio of 0.85 to 0.15, respectively. The test data was isolated to guarantee that there was no information leakage between the train and test datasets. Yet, the abovementioned preprocesses (median imputation and normalization) had to be executed for the test data set too, to avoid dimension mismatches and scale impropriety. Therefore, the same preprocesses were applied to the test data set separately.

2.4.2. Machine Learning Algorithms

For the classification of the aforementioned foot conditions (see Table 1) 5 conventional ML algorithms were evaluated in multiclass classification strategy and their results were compared: Support Vector

Machines (SVM) (26), Random Forest (RF) (27), Logistic Regression (LREGR) (28), K-nearest Neighbor (KNN) (29), and a majority voting (MV) (25) algorithm were trained separately. The MV model was implemented with a weighted soft voting technique that uses the trained SVM, KNN, RF, and LREGR models and gives the weighted voting outcome of them as the final output.

2.4.3. Hyper-parameter Tuning

Within a 3-times repeated 10-fold cross-validation, SVM, KNN, and RF models were tuned with a randomized search algorithm before training for finding the best hyper-parameters (30). For the LREGR model, a computationally expensive grid search algorithm was required for ensuring model convergence. Table 3 lists the tuned hyper-parameters of the first four models. MV weights were optimized through a randomized search algorithm after training the first four models (SVM, KNN, RF, LREGR).

Table 3
Hyper-parameters of the SVM, RF, LREGR, and KNN models

Model	List of hyper-parameters
SVM	Type of kernel, kernel coefficient (gamma), regularization parameter
RF	The number of trees in the forest, split quality criterion, the minimum required samples for splitting a node, the minimum required samples for being a leaf node, maximum depth of trees, the maximum number of features for splitting, the existence of bootstrapping
LREGR	The regularization strength, the type of solver algorithm, the maximum number of iterations
KNN	The type of distance metric, the type of weight function, the type of algorithm for computing nearest neighbor, the size of leaf (for the requiring algorithms)

Further details about the mentioned hyper-parameters of the models can be found in the documentation of Sklearn Framework for (25).

2.4.4. Training and Testing

The tuned models were trained and fitted to the training dataset, which consists of the selected scalar features of the functional angles (as inputs) and their foot conditions (as outputs). 90% of the train data was used for training and 10% for validation. 3-times repeated 10-fold cross-validation was applied. While the first 4 models do not need a special order of training, the majority voting model has to be trained at last, since it uses the outputs of the first 4 models as inputs.

Multiclass averages of accuracy, recall, and F1 scores of the models were calculated for quantifying the classification success of the models. Counts of true positives, true negatives, false negatives, and false positives of all foot conditions were taken into account for calculating these scores according to Formulas in Table 4, where T, F, P, and N stand for True, False, Positive, and Negative. Following these Formulas, the specific use of the terms accuracy, recall, and F1 score was described (31). The highness of these performance measures altogether, without them being correlated, indicates the success of the classification (32). These scores were calculated for the isolated test dataset.

Table 4
Performance measures, their formulas, and definitions (31)

Measure	Formula	Description
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Average per-foot condition effectiveness of a classifier model
Recall	$\frac{TP}{TP+FN}$	Effectiveness of a classifier model to identify foot conditions if calculated from sums of per-subject decisions
Precision	$\frac{TP}{TP+FP}$	An average per-foot condition agreement of the subjects' real foot conditions with those predicted by the classifier model
F1 Score	$2 * \frac{Precision*Recall}{Precision+Recall}$	Relations between data's positive labels and those given by the classifier model, based on sums of per-foot condition decisions

2.4.5. Model Interpretation

Before model interpretation, the features were recalculated to reach their original values by applying the inverse of normalization and skewing. The input features of the trained models were evaluated in terms of their relation to the decisions of the models. For this purpose, a model agnostic interpretation algorithm, namely "LIME" (Local Interpretable Model Explanation) algorithm was used, which observes the effect of each input feature on the output by doing perturbations on the inputs (7). For each functional foot condition and each model, this algorithm was executed with 20 random subjects. The number of selections of each input feature by LIME algorithm was noted. The features that were selected mostly by LIME were evaluated as the most related features for classifying the corresponding specific foot condition.

3. Results

3.1. The tuned hyperparameters

The tuned hyperparameters of the aforementioned models are as follows.

In the SVM a linear kernel with a kernel coefficient (gamma) and the regularization parameter (C) of 0.05 and 70, were used respectively.

In the RF, the number of trees in the forest was selected as 6984, the GINI impurity criterion was used, the minimum required subjects for splitting a node was selected as 2, and the minimum required subjects for being a leaf node was selected as 1, max depth of trees was limited to 40, the square root of the total feature number was selected as the maximum feature number to be considered and a bootstrapping was implemented.

In the LREGR, the inverse of regularization strength (C) was selected as 0.7 and the sag algorithm was executed with a maximum iteration number of 5000 for converging the model.

In the KNN, the Manhattan distance metric was used for 6 neighbors with uniformed weights and a KDTree algorithm was used for nearest neighbor calculation, for which a leaf size of 10 was used.

3.2. Training and Classification Scores

The learning curves and scalability graphs are shown in Fig. 3 for SVM, RF, LREGR, KNN, and MV respectively. The faded areas show the standard deviations in each. The learning curve graphs show the average accuracy score vs training examples for training and cross-validation sets. During the training, the cross-validation scores converged towards the training score, however not entirely. The gap between the training and cross-validation curves indicates a little overfitting that may decrease as the data size becomes bigger with the new data. In the scalability graphs, the linear trends along with the change in the number of data indicate the robustness of the models.

The average test scores for each model are shown in Table 5 below. All of the trained models achieved high scores in every metric, meaning that not only the foot conditions were predicted correctly, but also false prediction was too rare.

Table 5
Test scores for each model

Trained Model	Mean Accuracy	Mean Recall	Mean F1 Score
SVM	0.83	0.83	0.82
LREGR	0.83	0.90	0.83
RF	0.83	0.87	0.83
KNN	0.86	0.97	0.86
MV	0.86	0.90	0.86

Additionally, we split the train and test data with a ratio of 0.75 and 0.25, respectively and similar scores have been reached. This shows that the trained models are sufficiently robust.

3.3. Model Interpretation Outcomes

The most related features for each foot condition are shown in Fig. 4. In this Figure, the most selected kinematic features as relevant by LIME algorithm are listed for each foot condition.

In each feature name in Fig. 4, the following information is to be read: The short name of the functional angle the feature belongs to (as in Table 1), the type of the derived time series if not itself, the segment of the gait cycle to which the feature belongs to. If the standard deviation time series is used instead of the mean time series, it is indicated with a |STD| in the short time series name. The length of the bars corresponds to the number of selections of a feature as one of the most dominant ones.

Reading example of Fig. 4: In the classification of Tibiotalar osteoarthritis + partial ankle replacement, minimum medial arch angle during both stance and swing phases (the longest bars in the figure) were

found as the most relevant feature by RF, KNN, and MV models. The other relevant features are to be seen in the figure.

4. Discussion

In this study, it is aimed to automate the feature selection and foot condition identification to facilitate the diagnosis of foot conditions and hence, to support experts and physicians in their clinical assessment. An ML pipeline is achieved using the state-of-the-art feature elimination and ML algorithms. LIME is utilized to interpret the outcomes of the ML models so that the experts using this pipeline have feedback about how ML models predicts the foot conditions.

Our findings indicate that all the algorithms showed good prediction with a minimum of 0.82 F-1 scores. The success scores of the algorithms are similar but KNN and MV yield the highest scores. However, since MV utilizes all, it will likely have the highest score when there is new data.

When the dataset has a class imbalance, namely some classes have very low data compared to the others, the ML models in general could tend to learn more about the classes with high data. RF is known to overcome class imbalance inherently, however, the test scores in Table 5 indicate that other models also perform well to deal with the class imbalance.

High success scores indicate mathematically strong relationship between the selected features and foot conditions. Their physiological relevance needs to be medically confirmed with the physicians. Moreover, different mathematical features can be computed manually and easily added to the created models. Additionally, automatic feature generation or even featureless classification methods can be explored.

The ML methods generate black-box models, meaning that how these models make a prediction is not transparent and very difficult to interpret. It is a common problem in neural network-based models, as well and there are new methods to interpret the resulting predictions of the models. In this study, LIME was used for the interpretation of the model predictions by showing the dominant (i.e. most contributing) features in the prediction. Furthermore, if the same features are selected by most of the models, it can be stated that those features might have strong relevance in the classification. To understand the clinical relevance, the LIME results will be discussed with the experts and physicians for the validation of the gained insights. The LIME results may help them discover unknown features that might be critical in a successful inspection of the foot conditions.

Note that the data size is comparably small (i.e. 348 subject data). The models will improve, as the new data arrives or new features are added. Furthermore, the model can be extended to classify other functional foot conditions that are not included in this study. Progress in digital patient records of specialized treatment centers or even through disease registries will facilitate and strengthen the assistance opportunities of ML and xAI in orthopedic diagnostic decision-making. The trained models do not require high computational resources for making the classification. Therefore, they can be easily integrated into any type of software (e.g. web-based) in the future.

5. Conclusions

We build a ML pipeline that has feature selection steps and includes 5 classification methods and an explainable AI part. In the feature selection part, we extract features manually and eliminate the ones with low variance. We compared the ML models and showed their ability in the classification of the foot conditions that might reduce clinicians' effort on time-consuming evaluation of the gait data. The findings of the Explainable AI method (i.e. LIME) indicate that it can help to understand the reason behind the classification results, that might give insight in mining new knowledge.

Abbreviations

AI
Artificial Intelligence
CP
Cerebral Palsy
HFMM
Heidelberg Foot Measurement Method
KNN
K-nearest Neighbor
LDA
Linear Discriminant Analysis
LIME
Local Interpretable Model-agnostic Explanation
LREG
Logistic Regression
LRP
Layer-Wise Relevance Propagation
ML
Machine Learning
MV
Majority Voting
PCA
Principal Component Analysis
RF
Random Forest
SHAP
Shapley Additive Explanations
Std
Standard Deviation
SVM

Support Vector Machine
xAI
Explainable Artificial Intelligence

Declarations

Ethics approval and consent to participate

Ethics approval was obtained from Ethical Commission of the Medical Faculty of the University of Heidelberg “Ethikkommission der Med. Fakultät der Universität Heidelberg“ S-850/2019.

Consent to participate: Contacting patients and asking for their consent for a retrospective analysis of their data for the described research purpose is not done in this study. The number of persons to be contacted would be large, and in some cases the contact would be in relation to medical care that took place several years ago, which means that a corresponding change of address of the persons concerned is not unlikely. The protection of the anonymized biomechanical data is of secondary importance compared to the described research interest, because in particular

a) the data to be evaluated are already available at the research centre and the original collection took place in the context of routine medical care and the data are to be processed here only for research purposes, and

b) for the purpose of the research, only persons who were already authorized to inspect personal data on the occasion of routine medical care are allowed to do so, and

c) Personal data will not be passed on to external bodies.

Consent for publication

Not applicable

Availability of data and materials

We have the approval to use these data retrospectively for this specific study. We may not give these data to any third party without formulating a specific scientific goal (like we did here in our collaboration). Also they are not available upon simple request to the corresponding author but only via a regular study.

Competing interests

This study is conducted without any commercial relationship and therefore the authors declare that there is no conflict of interests or competing interests to disclose.

Funding

This research has been funded by the Federal Ministry of Education and Research of Germany (Bundesministerium für Bildung und Forschung; BMBF) within the project “3DFOOT” (project number 01EC1907D).

Authors' contributions

MEÖ:

- Designed and implemented the Machine Learning and explainable AI workflow
- Wrote Methods and Results sections
- Prepared figures 2-4 and all tables

AY:

- Made the initial implementation of the data processing pipeline and supervised the Machine Learning and explainable AI workflow
- Wrote Abstract, Introduction and Discussion sections
- Prepared figure 1

MEÖ and AY contributed equally.

FS and SC: Collected data and prepared it for further processing.

SW: Supervised in the perspective of practical usage of gait analysis as well as about the clinical applicability of explainable AI.

US: Supervised in the perspective of foot Biomechanics.

All authors reviewed the manuscript.

Acknowledgements

Not applicable

References

1. Halilaj E, Rajagopal A, Fiterau M, Hicks JL, Hastie TJ, Delp SL. Machine learning in human movement biomechanics: Best practices, common pitfalls, and new opportunities. Journal of Biomechanics. 2018;81:1-11.
2. Deluzio KJ, Wyss UP, Costigan PA, Sorbie C, Zee B. Gait assessment in unicompartmental knee arthroplasty patients: Principal component modelling of gait waveforms and clinical status. Human

Movement Science. 1999;18(5):701-11.
3. Chen Y, Elenee Argentinis JD, Weber G. IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research. Clinical Therapeutics. 2016;38(4):688-701.
4. Isci S, Kalender DSY, Bayraktar F, Yaman A. Machine Learning Models for Classification of Cushing's Syndrome Using Retrospective Data. IEEE J Biomed Health Inform. 2021;25(8):3153-62.
5. Borjali A, Chen AF, Muratoglu O, Morid MA, Varadarajan K, editors. Deep Learning in Orthopedics: How Do We Build Trust in the Machine?2020.
6. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
7. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier2016. 97-101 p.
8. Simon J, Doederlein L, McIntosh AS, Metaxiotis D, Bock HG, Wolf SI. The Heidelberg foot measurement method: Development, description and assessment. Gait & Posture. 2006;23(4):411-24.
9. Wolf S, Loose T, Schablowski M, Döderlein L, Rupp R, Gerner HJ, et al. Automated feature assessment in instrumented gait analysis. Gait Posture. 2006;23(3):331-8.
10. Schöllhorn WI. Applications of artificial neural nets in clinical biomechanics. Clinical Biomechanics. 2004;19(9):876-98.
11. Figueiredo J, Santos CP, Moreno JC. Automatic recognition of gait patterns in human motor disorders using machine learning: A review. Medical Engineering & Physics. 2018;53:1-12.
12. Van Gestel L, De Laet T, Di Lello E, Bruyninckx H, Molenaers G, Van Campenhout A, et al. Probabilistic gait classification in children with cerebral palsy: A Bayesian approach. Research in Developmental Disabilities. 2011;32(6):2542-52.
13. Nüesch C, Valderrabano V, Huber C, von Tscharner V, Pagenstert G. Gait patterns of asymmetric ankle osteoarthritis patients. Clin Biomech (Bristol, Avon). 2012;27(6):613-8.
14. Alaqtash M, Sarkodie-Gyan T, Yu H, Fuentes O, Brower R, Abdelgawad A, editors. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society; 2011 30 Aug.-3 Sept. 2011.
15. Horst F, Lapuschkin S, Samek W, Müller KR, Schöllhorn WI. Explaining the unique nature of individual gait patterns with deep learning. Sci Rep. 2019;9(1):2391.
16. Slijepcevic D, Horst F, Lapuschkin S, Horsak B, Raberger A-M, Kranzl A, et al. Explaining Machine Learning Models for Clinical Gait Analysis. ACM Trans Comput Healthcare. 2021;3(2):Article 14.
17. Döderlein L, Häfner R, Wenz W, Schneider U. Fussdeformitäten: Der Spitzfuss/Der Hackenfuss: Springer Berlin Heidelberg; 2013.

18. Caroll N, Döderlein L, Wenz W, Rauschmann MA, Schneider U. FussdeformitÄten: Der Knickplattfuss: Springer Berlin Heidelberg; 2013.
19. Döderlein L, Fixsen JA, Wenz W, Schneider U. Der Klumpfuss: Erscheinungsformen und Behandlungsprinzipien jeden Alters. Differentialdiagnose und Differentialtherapie: Springer Berlin Heidelberg; 2013.
20. Döderlein L, Wenz W, Schneider U. Der Hohlfuß. Fussdeformitäten: Der Hohlfuss. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001. p. 1-107.
21. Döderlein L, Wenz W, Schneider U. Der Ballenhohlfuß. Fussdeformitäten: Der Hohlfuss. Berlin, Heidelberg: Springer Berlin Heidelberg; 2001. p. 109-75.
22. Engels JM, Diehr P. Imputation of missing longitudinal data: a comparison of methods. J Clin Epidemiol. 2003;56(10):968-76.
23. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37; Lille, France: JMLR.org; 2015. p. 448–56.
24. D'Amour A, Heller K, Moldovan D, Adlam B, Alipanahi B, Beutel A, et al. Underspecification Presents Challenges for Credibility in Modern Machine Learning2020.
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(null):2825–30.
26. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20(3):273-97.
27. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32.
28. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature:: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology. 2001;54(10):979-85.
29. Altman NS. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician. 1992;46:175-85.
30. Paper D. Scikit-Learn Classifier Tuning from Simple Training Sets. Hands-on Scikit-Learn for Machine Learning Applications: Data Science Fundamentals with Python. Berkeley, CA: Apress; 2020. p. 137-63.
31. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009;45(4):427-37.
32. Seliya N, Khoshgoftaar TM, Hulse JV, editors. A Study on the Relationships of Classifier Performance Metrics. 2009 21st IEEE International Conference on Tools with Artificial Intelligence; 2009 2-4 Nov. 2009.

Figures

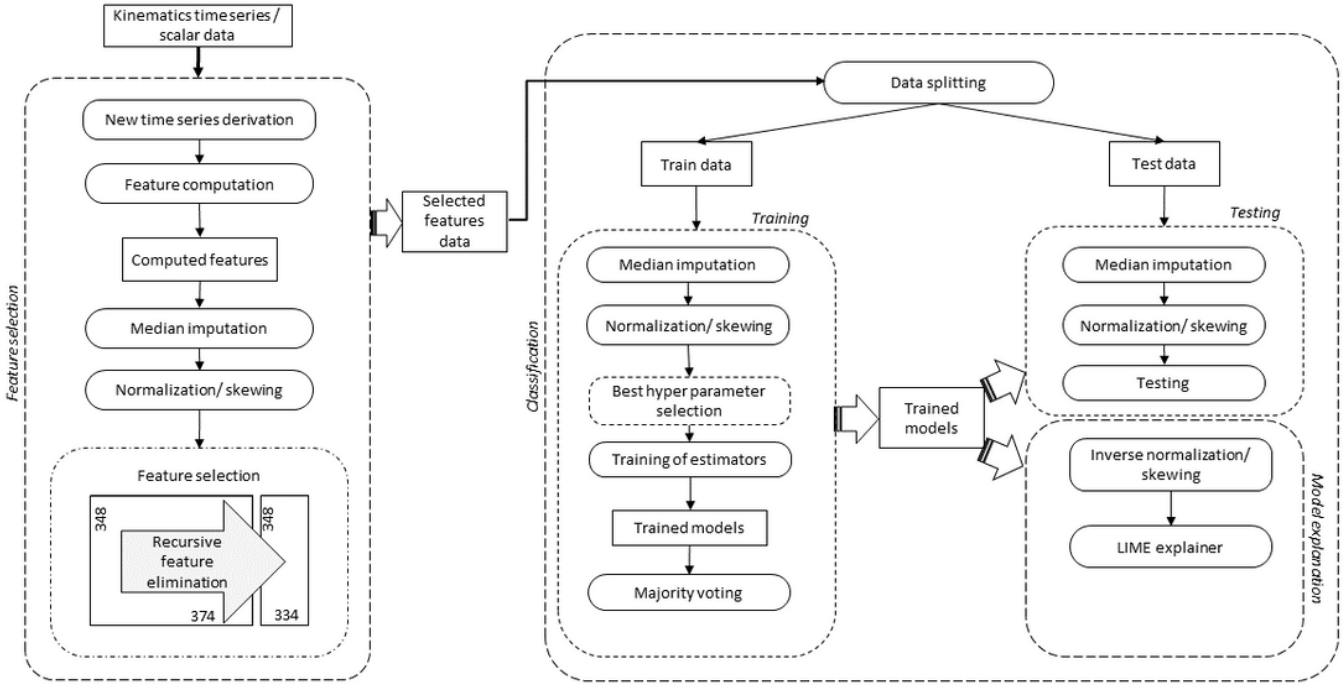

Figure 1

Overall method flowchart

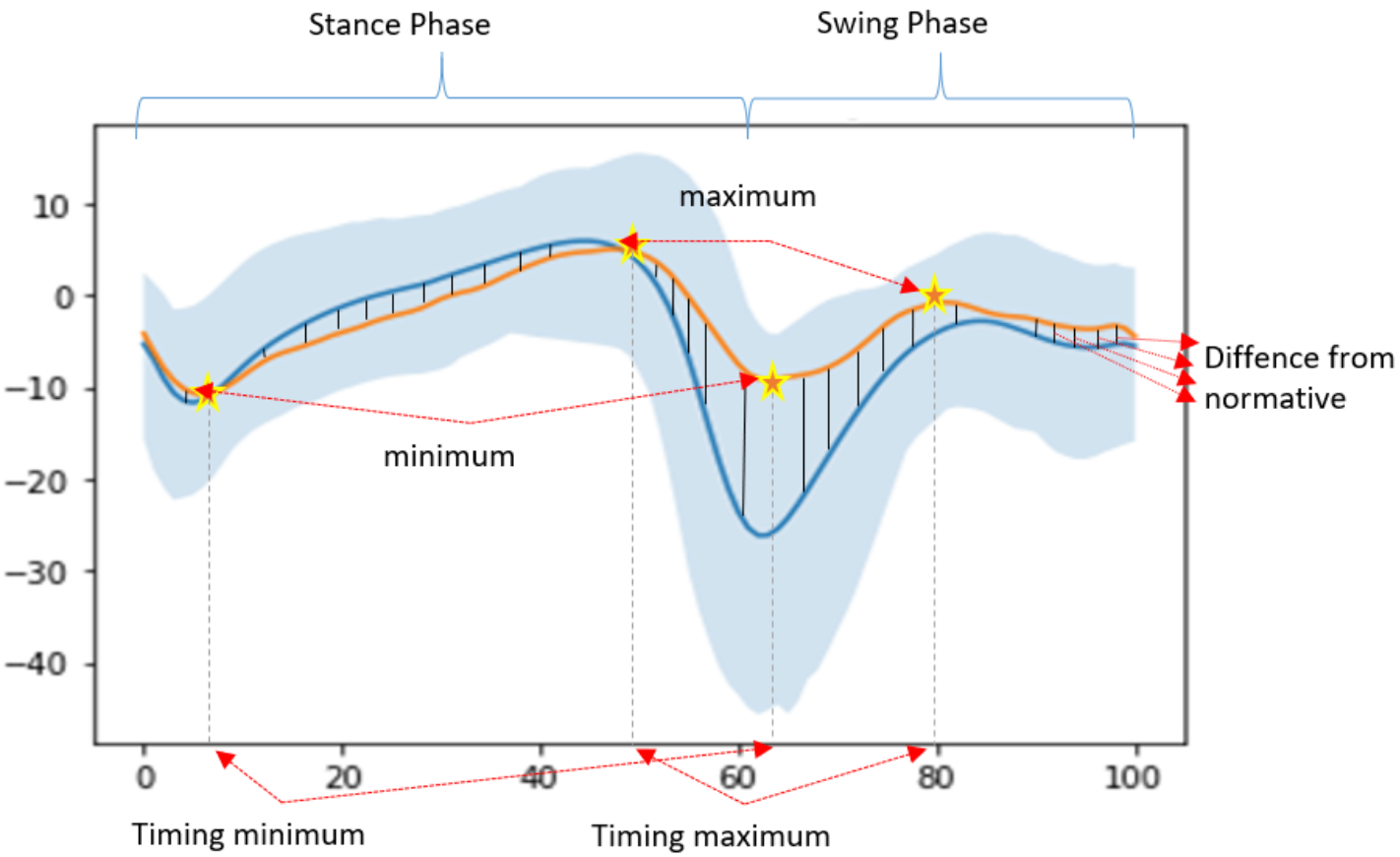

Figure 2

Computed features of an example Time Series. The orange line shows the tibia dorsi flexion of the left foot for a subject with hallux rigidus. The blue line and band show the mean value and variation of the reference normal time series.

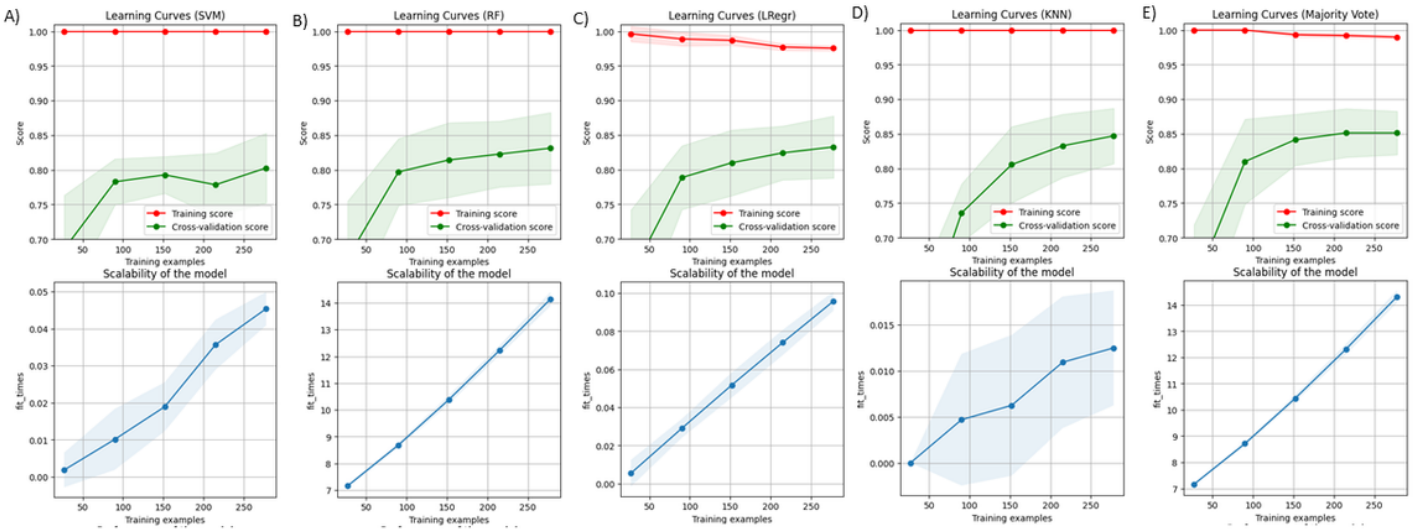

Figure 3

Training graphs of the models. The learning curves and scalability of the model graphs for A) SVM, B) RF, C) LREGR, D) KNN, E) MV

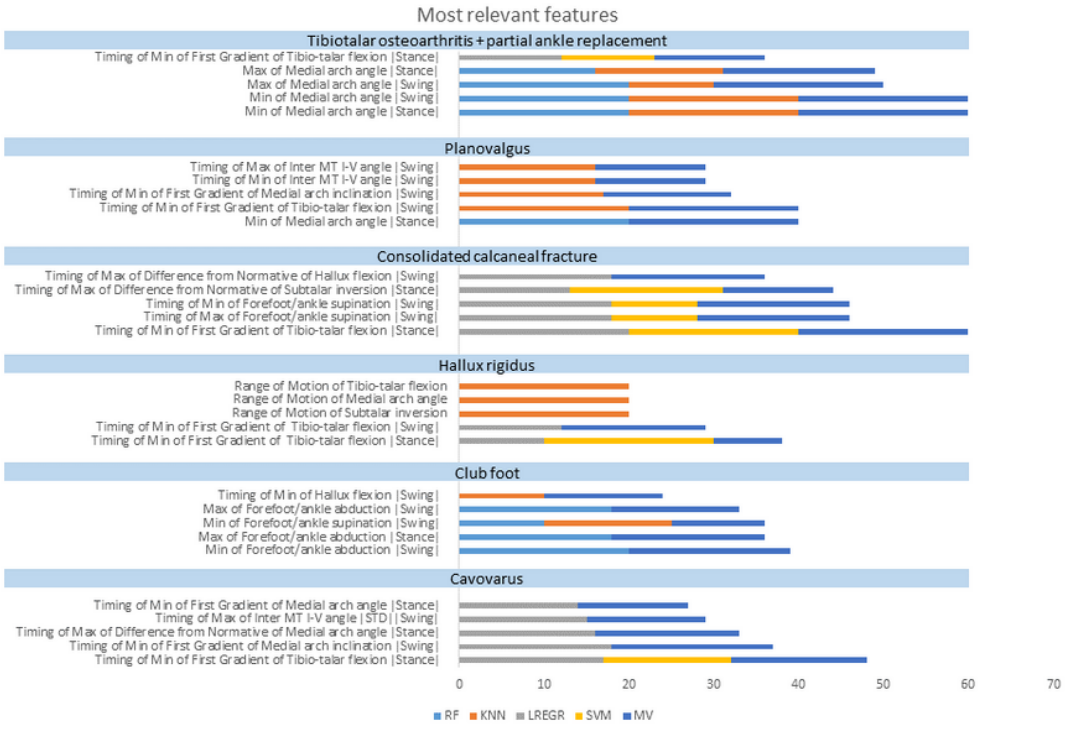

Figure 4

Classifier Features for each foot condition. Colors representing the models: Green, dark blue, orange, light blue, and yellow for RF, KNN, SVM, and MV, respectively. The bar length is related to the number of subject data on which the feature is dominant.