EXAMINING MACHINE LEARNING CLASSIFICATIONS WITH EXPLAINABLE AI AIDS INTERPRETATION OF WRIST BIOMECHANICS

Isaly Tappan[1], Erica M. Lindbeck[1], Jennifer A. Nichols[2], and Joel B. Harley[1]
[1]Department of Electrical and Computer Engineering, University of Florida
[2]J. Crayton Pruitt Family Department of Biomedical Engineering, University of Florida
email: i.tappan@ufl.edu

Introduction

The last decade has seen advances in the speed, reliability, and accuracy of biomechanical data collection. As datasets increase in size and complexity, biomechanists have turned to artificial intelligence (AI) to aid their analyses. However, a significant drawback of using AI is its black-box character, which obfuscates the reason for any given prediction. To alleviate this, explainable AI (XAI) has evolved to elucidate complex relationships between input data and the decisions of an AI system. Here, we explore whether local explanations (i.e., explaining what features best explain individual classifications) can enhance the interpretability of biomechanics data derived from musculoskeletal simulations. We use XAI to explain how a machine learning algorithm classifies simulated lateral pinch data as belonging to either healthy or types of surgically altered wrists. This simulation-based classification task is analogous to using biomechanical movement and force data to clinically diagnose a pathological state.

Methods

A simulation dataset was generated using published models and methods [1] in OpenSim. Briefly, three models were created to represent the healthy wrist and two surgeries for wrist osteoarthritis (PRC: proximal row carpectomy and SE4CF: scaphoid-excision four-corner fusion). Simulations of lateral pinch (n = 315) were generated by varying the anthropometric scaling (90-110% of baseline) and target thumb-tip endpoint force (10-50N in increments of 10N). Each simulation can be thought of as data from a synthetic patient. Simulation outputs included a 3D thumb-tip endpoint force and six joint angles across the wrist and thumb for a total of nine time-varying features. A random forest classifier [2] composed of 100 trees was trained to predict wrist impairment using 15-fold cross-validation (93% training simulations and 7% testing simulations).

Within every unique validation fold, an XAI framework known as Local Interpretable Model-Agnostic Explanations (LIME) [3] was applied to each time point in the testing set, thereby producing nine feature importance scores per time point per subject. The feature importance score indicates how much the prediction probability of the top class would change if the given feature were perturbed, with negative values indicating the feature contributed toward a class other than the top prediction. Following cross-validation, the importance scores were averaged across time to produce nine final feature importance values (i.e., scores defined for each synthetic subject). The final values were subsequently averaged across each classification to produce explanations specific to each impairment (nonimpaired, PRC, SE4CF). Final values were also averaged across musculoskeletal model sizes (produces explanations for combinations of target force and impairment classification), and across target forces (produces explanations for combinations of model size and impairment classifications). From these averages, we can evaluate whether explanations for identifying impairment change based on the characteristics (size and strength) of the synthetic patients.

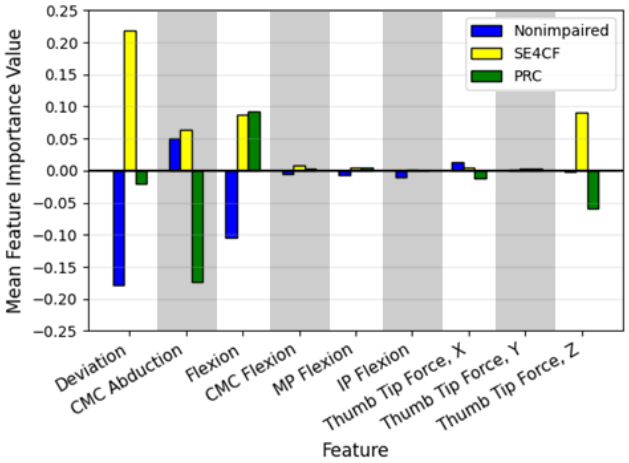

Figure 1: Final feature importance values averaged across impairment classification.

Results and Discussion

Our results demonstrate that LIME can be used to identify and explain the biomechanical features helpful for distinguishing nonimpaired and impaired synthetic patients. For example, the magnitude of the feature importance value for wrist flexion is large for all three conditions, but it is only negative in the nonimpaired condition (Fig. 1). This suggests that wrist flexion aids the classifier in distinguishing between nonimpaired and impaired states. Following this classification, the results show that a combination of radial-ulnar deviation and carpometacarpal (CMC) abduction can be used to further distinguish SE4CF versus PRC. This is illustrated by the difference in sign between the SE4CF and PRC importance values for radial-ulnar deviation and CMC abduction. Mean importance values across model sizes and target forces (not shown) illustrate the same trends, indicating these factors do not substantially influence which features are important. In addition, the explanations highlighted by LIME align with experimental data demonstrating that wrist flexion and radial-ulnar deviation are important for distinguishing these conditions [4].

Significance

This work demonstrates that XAI can identify the most important features for classifying impairments and how these features change with synthetic patient size and strength. Thus, XAI can effectively remove the black-box character of AI when used in biomechanical analyses. This capability could aid in elucidating the biomechanical mechanisms underlying impairment.

Acknowledgments

Funding from NIH NIBIB Trailblazer Award (R21EB030068).

References

[1] Nichols et al., 2017. *J Biomech*. 58.
[2] Pedregosa et al., 2011. JMLR 12: 2825-2830.
[3] Ribeiro et al., 2016. ACM SIGKDD: 1135-1144.
[4] Nichols et al., 2015. *J. Biomech*. 30.