One-Sided Classification for Interpretable Predictions

Jeroen Vuurens[a], Tony Andrioli[a], Erwin de Vlugt[a]

[a]*The Hague University of Applied Science, Lectorate of Health and Innovation*

Abstract

Many applications require machine learned models to be interpretable. This study explored potential pitfalls for learning a classifier: false identifiers, untraceable constructs and meaningless boundaries. Interpretable and acceptable results can be obtained by learning one-sided classifiers that substantiate the predicted class with a decisive argument. These one-sided classifiers can be combined in an OR-ensemble, which preserves the interpretability of the model as a whole. This is demonstrated on a set of kinematic data that was collected for the analysis of shoulder injuries, for which it provides equal or better results than state-of-the-art ensemble algorithms.

Keywords: Classification, Interpretability, Explainability, Diagnose, Kinematic recording

1. Introduction

A problem with existing machine learning algorithms is that it is difficult to explain the prediction in terms that are easy to interpret by humans [2]. This is especially true in the medical domain where medical staff needs to be able to trust the outcome [1]. Often, the results of machine learning algorithms are combined with other information sources to come to a diagnosis. For that purpose it is important that classifiers explain which observation led to the predicted class and that the decision is understandable and convincing.

Learning an interpretable model is challenging. Popular approaches learn either a univariate [11] or a rule-based model [7]. However, the learned decision rules may not be convincing when they deviate from accepted medical practice. Additionally, the way in which decision rules are combined in [11] may not be comprehensible. In this study, we analyze potential pitfalls in constructing an interpretable classifier. We argue that the results of a model should be acceptable to its user, and propose a partial solution in constructing one-sided classifiers that deliver a decisive argument for one class. An OR-ensemble can be used to combine multiple one-sided classifiers without sacrificing its interpretability.

This study uses a dataset of kinematic recordings of patients with rotator cuff disorders and that of a control group. One-sided classifiers are trained and then combined in an OR ensemble. The results are easy to interpret in practice and we show that this approach provides equal or better results than state-of-the-art machine learning models.

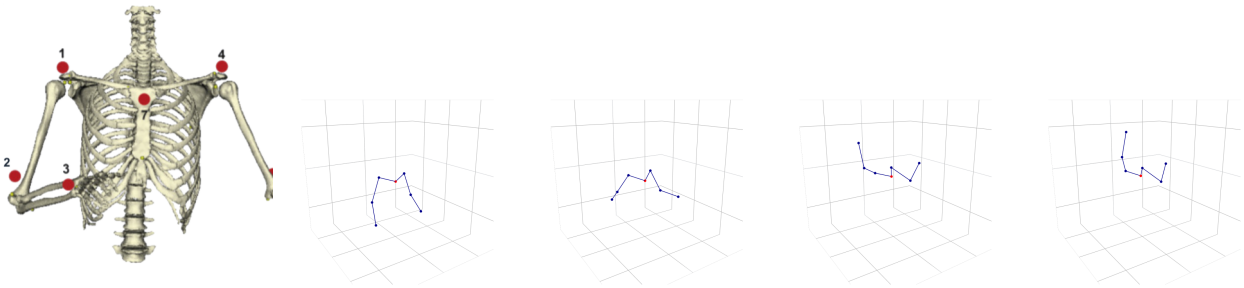
The remainder of this paper is structured as follows. Section 2 describes the dataset and the features that are used for learning. In Section 3 we analyze the results of logistic regression and show potential pitfalls in learning an interpretable classifier. We then describe a partial solution, by learning one-sided classifiers and combining these in an OR-ensemble in Section 4. We compared the results to the state-of-the-art machine learning algorithms and discuss the results in Section 5 and finally give the conclusion in Section 6.

2. Data

2.1. Sensor recordings

This study uses kinematic data of patients performing shoulder exercises in a medical facility. These descriptive recordings are used to find associations between the type of disorder and the arm-shoulder movements [6] [4] [5]. The

Email addresses: j.b.p.vuurens@hhs.nl (Jeroen Vuurens), a.andrioli@hhs.nl (Tony Andrioli), e.devlugt@hhs.nl (Erwin de Vlugt)



(a) (b)

Figure 1: Illustration of (a) how sensors are placed on the body, and (b) a sequence of sensor recordings for an abduction exercise, frontal view.

recordings consist of four exercises: Abduction (AB) where arms are simultaneously raised sideways from a hanging position to point upwards, Anteflexion (AF) where arms are raised to the front, Retroflexion (RF) where arms are pointed backwards as far as possible, and Exo- and Endorotation Low (EL) where the upper arm is pointed down and the lower arm is pointed forward, and then the shoulder is rotated to point the lower arms sideways while keeping the upper arm in place. Figure 1b illustrates with a stick diagram what such an abduction exercise looks like.

For the recordings, a 'Flock of birds'-sensor is attached to a participant to record a time-series of rotational angles of the bones while performing the exercises. The recordings were taken as follows: a specialist places sensors at the end of the bones of the shoulder, elbow, wrist and thorax (Fig 1a); the sensors are calibrated at the beginning of every recording; then the patient is instructed to do an exercise, during which a time-series of 3x3 rotation matrices is recorded for every sensor, describing its rotation. Each of the sensors measure their rotation towards a sensor that is placed in a fixed position on the floor behind the patient. The recorded 3D-Euclidian coordinates are used in this study.

2.2. Labels

In the dataset, 36 participants are labeled as patients (having severe rotator cuff ruptures) and 29 participants are labeled as members of the control group (having no rotator cuff injury). As is exemplary for many classes in the medical field, the patient class consists of several subclasses, e.g. as having problems in their left or right shoulder, problems in the abduction, anteflexion or retroflexion planes. However, there are no labels for these. We excluded two participants that did not complete all four exercises.

2.3. Features

Since the dataset is small (n=65) and high-dimensional (4 exercises for each participant, on average 160 frames, 7 sensors), we needed to extract a small set of features or representations that are meaningful for classification to avoid overfitting. In practice, physicians diagnose rotator cuff injuries in various ways, and one way that is closely related to the data at hand is to observe how the shoulder blades move in the lower region (upper arms hanging down). However, in this case the 3D Euclidian coordinates simply do not contain the information to extract that observation. Instead, we use the minimum and maximum shoulder angles for the Abduction (AB), Retroflexion (RF) and Anteflexion (AF) exercises and the minimum and maximum angle of the lower arm for the Endo- and Exo-rotation Low exercise (EL). The shoulder angle is computed between a vector that is drawn through the two shoulders (sensor 1 and 4 in Fig 1a) and a vector through the shoulder and the elbow of respectively the right or left arm (e.g. for the right shoulder angle the vectors through sensor 1 and 2). The maximum angle is inspired by the inability of patients with frozen shoulder injuries to raise their arm beyond some angle, and the minimum angle because physicians primarily observe movements in the lower region.

3. Analysis

In this Section, we analyze logistic regression models and discuss pitfalls in classifier design and how interpretability of a model is compromised. We then propose to tweak the training set in order to construct interpretable classifiers

features	recall	precision	accuracy
all_features	0.917	0.943	0.923
all_left	0.889	0.889	0.877
all_max	0.833	0.857	0.831
all_AB	0.806	0.879	0.831
all_right	0.806	0.879	0.831
all_EL	0.861	0.795	0.800
all_AF	0.778	0.800	0.769
all_min	0.750	0.818	0.769
max_AF	0.750	0.771	0.738
max_AB	0.722	0.765	0.723
min_AB	0.722	0.743	0.708
max_EL	0.750	0.711	0.692
min_EL	0.722	0.722	0.692
min_AF	0.694	0.595	0.569
min_RF	0.694	0.521	0.477
max_RF	0.722	0.510	0.462
all_RF	0.611	0.478	0.415

(a) Multivariate

features	recall	precision	accuracy
left_max_AF	0.750	0.771	0.738
left_max_EL	0.750	0.711	0.692
right_min_AB	0.778	0.700	0.692
left_min_EL	0.722	0.703	0.677
left_min_AB	0.722	0.703	0.677
left_max_AB	0.667	0.706	0.662
right_max_AB	0.639	0.697	0.646
right_min_EL	0.667	0.667	0.631
right_max_EL	0.667	0.649	0.615
right_min_AF	0.778	0.609	0.600
right_max_AF	0.611	0.647	0.600
right_min_RF	0.806	0.569	0.554
right_max_RF	0.944	0.540	0.523
left_min_RF	0.917	0.532	0.508
left_max_RF	0.778	0.538	0.508
left_min_AF	0.667	0.522	0.477

(b) Univariate

Table 1: Results for logistic regression models that predicts whether a person has a rotator cuff injury based on the given feature(s), e.g. right for right shoulder, max for maximum angle, AF in the Anteflexion exercise.

that can be externally validated.

3.1. Interpretability

Within the medical domain, the outcome of a classification model may affect decisions regarding a patient and therefore, should be presented in a way that excludes wrong interpretations. There are different opinions in literature regarding the requirements with respect to the interpretability and explainability of classification models. In some cases it may be sufficient to explain to what extent the features are important to reach a decision (e.g. [10]). Alternatively, post-hoc interpretability considers a classification model to be a black-box and attempts to add a readable explanation [2]. However, special care must be taken to avoid models that generate misleading explanations. [7] propose using sets of independent if-then decision rules, which are easy to interpret. Still, that is not a guarantee that the outcomes can be trusted.

3.2. False Indicators

In this study we fit all logistic regression models to maximize the log likelihood of Equation 1 and use a leave-one-out strategy to avoid overfitting. In Table 1a, we fit several multi-variate logistic regression models. We observe that the all_features model obtains an accuracy of 92%, however, these results will appear not be valid. The all_min model classifies 77% correctly, which is extraordinary when we consider that the decision is based on the participants' arms hanging down, which is not a position that people with shoulder injuries struggle with. When we explore this further by constructing separate univariate logistic regression models (Table 1b), it appears that the minimal angle of the right arm in the AB exercise explains 69% of the dataset. A visual inspection of the exercises reveals that a possible explanation is that half of the patients have their hands on their lap at the start of the exercise, while the control-group members do not. We consider this a False Indicator, since it is not likely that this generalizes to other situations. A model that incorporates a feature of which we do not know what it measures lacks interpretability.

$$\log \frac{p(y=1|x;\theta)}{p(y=0|x;\theta)} = \theta^T \cdot x \tag{1}$$

		left_max_AB	right_max_AB	left_max_AF	right_max_AF	left_max_EL	right_max_EL	left_max_RF
right_max_AB	tp	337, 321, 313						
left_max_AF	fn		337					
left_max_AF	tp	336, 337	336, 321, 337, 313					
right_max_AF	tp	337, 307, 313, 314, 303	321, 339, 307, 314	336, 313, 307, 303				
left_max_EL	tp	322, 307, 311, 327, 333, 318	321, 322, 325, 327, 330, 337, 339, 311, 318	337, 311, 314, 318, 303	337, 311, 325, 327, 330, 314, 333			
right_max_EL	tp	322, 327, 332, 333, 318, 303	321, 322, 327, 330, 332, 337, 339, 313	337, 332, 318, 303	322, 327, 330, 332, 333, 307, 313, 314	332, 325, 311		
left_max_RF	fn		321		303			
left_max_RF	tp	321, 324, 327, 331, 332, 303, 307, 314, 318	321, 325, 327, 330, 331, 333, 335, 337, 306, 313, 314	336, 321, 337, 331, 332, 318, 303	321, 325, 327, 330, 303, 335, 306, 313	321, 322, 331, 332, 333, 304, 339, 311, 314	321, 322, 325, 331, 333, 306, 307, 339, 314	
right_max_RF	fn		337		307			304, 315, 316
right_max_RF	tp	321, 327, 331, 332, 303, 337, 307, 311, 312, 314, 318	325, 327, 330, 331, 332, 333, 337, 306, 307, 308, 311, 313, 314, 318	321, 331, 332, 303, 336, 307, 311, 318	321, 325, 327, 330, 331, 332, 306, 307, 339, 309, 340, 311, 313, 318	321, 322, 331, 332, 333, 304, 337, 339, 312, 314	321, 322, 325, 331, 333, 305, 306, 307, 337, 339, 311, 314, 319	329, 334, 335, 304, 337, 311, 312, 315, 316

Table 2: Experiment to show the lack of robustness in training. The number in the cells indicate for which patients a model that was learned on two features (row + column) gives an unexplainable prediction, when compared to the predictions from the 2 separate univariate logistic regression models for resp. the row and column feature. The label tp indicates when one of the univariate models outputs a true positive and the joined model does not. The label fn indicates when both of the univariate models output a true positive and the joined model does not.

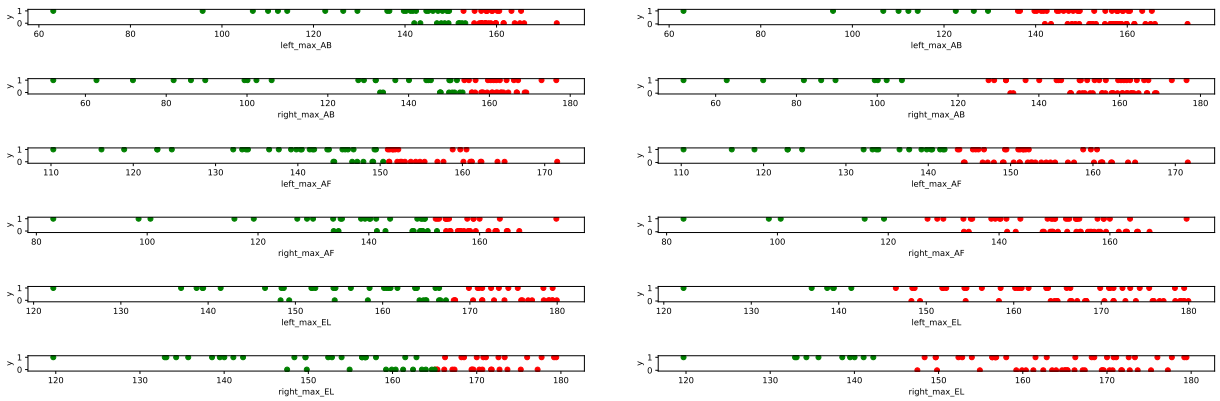
3.3. Untraceable constructs

Another potential pitfall is combining features without considering how they are related. In this dataset, there are features for left- and right shoulders. However, a person is often only injured in one shoulder. When we combine left and right shoulder features it may accumulate in a non-interpretable manner. This is demonstrated by comparing the predictions given by bivariate logistic regression models to the predictions from the two corresponding univariate logistic regression models. Table 2 lists inconsistencies between these models that can obstruct the interpretability. Specifically, sometimes a diagnosis made on a single feature correctly classifies a patient, but rather a bivariate model that has an additional feature does not. E.g. for patient 337, a correct diagnosis is given by models learned on either right_max_AB or right_max_RF, but a model learned on both features fails to correctly classify this patient. (type 2) for patient 336 the univariate model that uses right_max_AB correctly identifies the patient, but the joined model that uses right_max_AB and left_max_AF does not. In other words, the information that a participant belongs to the patient class is there, but adding possibly irrelevant information leads to an incorrect result. Still, the additional feature may reveal something about the flexibility or age of a person. And this is a dilemma: when we cannot tell if the information is accumulated in a valid manner, if a reasonable decision boundary was used, a model can no longer be trusted. For the described scenario, we argue that the interpretability of the model may be improved by specifically showing that a construct that is meaningful in the medical domain is being represented (e.g. age of flexibility), and how that representation is used in classification. There is interesting existing work on 'feature importance', e.g. LIME [9] and SHAP [8], but we note that their application may be limited to cases where both the features and their aggregation are evident.

3.4. Meaningless boundaries

Many applications require machine learned models to be interpretable. [7] proposes to bridge the gap between models and domain experts by sets of univariate if-then decision rules, since these are easy to understand. However, easy to understand is not the same as being acceptable. Classification algorithms find optimal decision boundaries to separate classes, but without any consideration whether or not these are acceptable to humans. To illustrate this, Fig 2a show how the patients (y=1) and control-group members (y=0) are distributed over single feature values. Leave-one-out logistic regression chooses a decision boundary that provides optimal classification results for both classes, as seen by the cross-over point between the red and green dots. However, the chosen boundary does not justify a positive or negative diagnosis.

[11] criticize the decision boundaries taken by [7] and argue that for univariate models one can often distinguish between an area where two classes are indistinguishable, and one or more area where one class is more likely. They propose an automated way to find a cutoff point. However, the computed cut-off points as they propose may still not be very convincing evidence for a class prediction by itself, and their proposed majority votes ensemble over multiple classifiers combines the outcomes without considering the strength of each outcome. For a medical application, it is difficult to state that having 10 votes that say a patient is healthy outweighs 8 votes that say the contrary. Instead,



(a) using all values

(b) using values that differ from the control group

Figure 2: distribution patients (y=1) and control_group (y=0) of fitted univariate logistic regression models over the feature used (x-axis). In green are the predicted patients, in red the predicted members of the control group

human interpretation is easier if the explanation is simplified to a decisive argument. A decisive argument could be that the feature value for a person is very unlikely to be seen within the control group. To illustrate this, we fit univariate logistic regression models on a training set from which we removed patients that have feature values that are similar to those observed in the control group (Fig 2b). Such classifiers are easy to comprehend and to validate by a domain expert, and can readily be used to explain why a positive label is given. These classifiers can be joined to construct a multivariate classifier, from which the explanation of a positive diagnosis is immediately clear, which we will describe further in the next Section.

4. Model

In this Section, we describe an approach to construct an interpretable classifier, by constructing one-sided logistic regression models and joining them in an OR ensemble.

4.1. One-sided logistic regression

In Section 3, it was shown that the decision boundary that is learned by logistic regression may not correspond to the decision boundary that is acceptable for domain experts (Fig 2a). This discrepancy is a form of bias. Although bias is commonly explained as an oversimplified model, justifying the addition of more features, this will not make the predictions more acceptable.

The main reason for the discrepancy in decision boundary is that the group of patients is heterogeneous. As illustrated in Table 2a, the distribution of the two classes over a feature reveals a large group for which we cannot distinguish between the two classes, and a small group of patients that have abnormal feature values. The acceptability can be increased by relaxing the default learning objective to try and identify all data points, but instead focusing one-sidedly on the 'identifiable' members of one class for that feature. We define 'identifiable members' in this study as the members that stand out from all members of the other class. Since in this case, patients are only likely to stand out only with a lower value, we are only interested in that one side. Equation 2 defines the set of training objects x_{train} we use to train a logistic regression classifier using members in the control group $x0$ and patients in $x1$ that have feature values that stand out according to a threshold τ. τ is computed using the lowest value for the feature observed in the control group $x0_{low}$, the mean observed feature value in the control group $\overline{x0}$, and a hyperparameter f to place the threshold τ slightly below $x0_{low}$. We can use f to control the trade-off between recall and precision, in practice $f = 1.1$ is a reasonable setting to maximize precision and thus to present a decision boundary that may be acceptable.

feature	recall	precision	accuracy
left_max_AF	0.528	1.000	0.738
right_max_AB	0.278	1.000	0.600
right_max_EL	0.278	1.000	0.600
left_max_AB	0.250	1.000	0.585
left_max_EL	0.167	1.000	0.538
right_max_AF	0.139	1.000	0.523

Table 3: One-sided logistic regression models for which only 'identifiable patients were used during training.

feature	recall	precision	accuracy
OR, f=1.1	0.722	1.000	0.846
Multivariate LR	0.806	0.829	0.800
AdaBoost LR	0.793	0.770	0.805
RandomForest	0.707	0.690	0.735
SVM	0.707	0.724	0.787

Table 4: An OR ensemble that combines the one-sided models compared to existing algorithms.

$$\begin{aligned} x0_{low} &= \arg\min_i x0_i \\ \tau &= (1-f)\cdot\overline{x0} + f\cdot x0_{low} \\ x_{train} &= x0 \cup \{a \in x1 | a < \tau\} \end{aligned} \tag{2}$$

Table 3 shows the results when learning one-sided logistic regression models using the training set according to Equation 2 using $f = 1.1$. The result is a set of high-precision one-sided classifiers. The external validity of these models can be ensured by checking the decision boundary of these models with domain experts or literature.

This approach resembles one-class classification, which labels outliers as positive. From the perspective of interpretability and acceptability, in a one-class classifier outliers are not well defined and the decision boundary is more difficult to understand. For example, in our case, a person that is more flexible than anyone else in the dataset would be an outlier, however, that is not indicative for the patient class.

4.2. OR-Ensemble

The one-sided classifiers in Table 3 are partial solutions. These can be combined by state-of-the-art ensemble algorithms, e.g. Boosting and Random Forests. However, similar to our analysis of multivariate logistic regression, this may lead to an incomprehensible algorithm. For example, it is unclear how the information from these features should be combined, since some results are mutually exclusive (e.g. results from the AB and EL classifiers), and some are mutually inclusive (e.g. left AB + left AF, right AB + right AF). Additional to this lack of interpretability, we will show that state-of-the-art ensembles do not perform well on this dataset.

Each one-sided classifier in Table 3 predicts a subset of patients with high precision. A logical way to join these is by using a Boolean OR-operator. This is analogous to the case where 3 physicians diagnose the same patient on 3 completely different pieces of evidence. If one of the physicians diagnoses a patient as positive, this should overrule a negative diagnosis from the physicians that did not observe that bit of information. A Boolean OR-operator makes the ensemble outcomes easy to explain or interpret, a positive label can be substantiated by the one-sided classifier that predicted a positive label, e.g. the maximum right arm elevation in de Abduction exercise was below 140° while non-patients typically are above 150°.

In Table 4, we show the results of a Boolean OR operator over the one-sided models in Table 3. We also compared the OR ensemble to state-of-the-art ensembles and observe that the accuracy and precision are better.

5. Discussion

The emphasis in this study is on the interpretability of the model. We focused on 6 fairly straightforward features from the subjects' kinetic recordings and observed that the decision boundaries learned by logistic regression cannot be trusted. By changing the focus to train only on identifiable patients, and by validating the decision boundary, the prediction of a positive diagnosis can be substantiated with a decisive argument. By combining multiple high-precision one-sided models in an OR-ensemble, state-of-the-art results can be obtained, while maintaining a model from which the outcomes are easy to interpret.

For diagnosing rotator cuff injuries, machine learned classifiers may be of limited use in practice (accuracy=0.846 in Table 4). An advantage of the proposed approach over existing models is that the predictions of the patient class are of high-precision (ideally precision=1) and the availability of readily interpretable information on the given diagnosis.

The recall of the model may be improved by adding one-sided classifiers over other features. The OR-ensemble may hold a few advantages over other ensemble algorithms. Since additional features are added as a one-sided classifier in an OR ensemble, there is less risk over overfitting, and OR-ensembles can operate with missing features. In this work, we focused on univariate models, because these allow to reason more easily on what a one-sided classifier would look like and why it is reasonable to expect such a classifier to produce a decisive argument. An interesting direction for future work is to explore how one-sided multivariate classifiers can be used. On this dataset, this would allow a comparison between the movements of both shoulders, to detect painful arcs (movements around pain), leaning towards the pain during an exercise, and slowing the tempo before painful areas, to name but a few. However, we should take care that such classifiers use an acceptable decision boundary, allowing them to deliver a decisive argument.

6. Conclusion

Within the medical domain, it is important to substantiate the given diagnosis with a valid explanation. In this paper we discussed common pitfalls when applying logistic regression without considering the interpretability of the resulting model. To address the problem of untraceable constructs and meaningless boundaries, a model is proposed were several one-sided classifiers are build by automatically removing patients from the training set that do not stand out from the control group. The resulting classifiers have a high precision and because of their simplicity, these classifiers that identify patients that stand out are easily validated by orthopedists and their outcome trusted as a valid reason for a positive diagnosis. Multiple of these univariate classifiers can then be combined using an OR-ensemble, to preserve the interpretability of the model as a whole. We show that for the diagnosis of rotator cuff injuries an OR-ensemble provides equal or better results than state-of-the-art ensemble algorithms. Additionally, an OR-ensemble over interpretable univariate classifiers can provide a straightforward explanation in case of a positive diagnosis.

References

[1] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[2] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[3] J. F. Henseler, A. Kolk, B. Zondag, J. Nagels, J. H. de Groot, and R. G. Nelissen. Three-dimensional shoulder motion after teres major or latissimus dorsi tendon transfer for posterosuperior rotator cuff tears. *Journal of Shoulder and Elbow Surgery*, 26(11):1955–1963, 2017.

[4] A. Kolk, P. B. de Witte, J. F. Henseler, E. W. van Zwet, E. R. van Arkel, P. van der Zwaal, R. G. Nelissen, and J. H. de Groot. Three-dimensional shoulder kinematics normalize after rotator cuff repair. *Journal of Shoulder and Elbow Surgery*, 25(6):881–889, 2016.

[5] A. Kolk, J. F. Henseler, P. B. de Witte, E. R. van Arkel, C. P. Visser, J. Nagels, R. G. Nelissen, and J. H. de Groot. Subacromial anaesthetics increase asymmetry of scapular kinematics in patients with subacromial pain syndrome. *Manual therapy*, 26:31–37, 2016.

[6] A. Kolk, J. F. Henseler, P. B. de Witte, E. W. van Zwet, P. van der Zwaal, C. P. Visser, J. Nagels, R. G. Nelissen, and J. H. de Groot. The effect of a rotator cuff tear and its size on three-dimensional shoulder motion. *Clinical Biomechanics*, 45:43–51, 2017.

[7] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684, 2016.

[8] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[10] W. Samek, T. Wiegand, and K.-R. Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[11] M. Sheth, A. Gerovitch, R. Welsch, and N. Markuzon. The univariate flagging algorithm (ufa): An interpretable approach for predictive modeling. *PloS one*, 14(10):e0223161, 2019.