Research Article

Comparison of feature importance measures as explanations for classification models

Mirka Saarela[1] · Susanne Jauhiainen[1]

Received: 10 July 2020 / Accepted: 4 January 2021

Abstract
Explainable artificial intelligence is an emerging research direction helping the user or developer of machine learning models understand why models behave the way they do. The most popular explanation technique is feature importance. However, there are several different approaches how feature importances are being measured, most notably global and local. In this study we compare different feature importance measures using both linear (logistic regression with L1 penalization) and non-linear (random forest) methods and local interpretable model-agnostic explanations on top of them. These methods are applied to two datasets from the medical domain, the openly available breast cancer data from the UCI Archive and a recently collected running injury data. Our results show that the most important features differ depending on the technique. We argue that a combination of several explanation techniques could provide more reliable and trustworthy results. In particular, local explanations should be used in the most critical cases such as false negatives.

Keywords Feature importance · Explainable artificial intelligence · Interpretable models · Random forest · Logistic regression

1 Introduction

Classification models have two main objectives [9]. First, they should perform well, meaning they should forecast the output for new given input features as accurately as possible. Second, they should be interpretable, that is, provide some understanding between the input features and the output. Usually, there is some tradeoff between these two objectives. For example, simple linear classification models are easy to understand and interpret but typically perform worse than non-linear models [10, 19, 24, 44], while complex prediction models with non-linear combinations of features tend to perform better (e.g., [32, 33, 41]) but are less interpretable. In other words, they often do a better job in classifying new instances correctly, but the reasons why a certain classification was made is hidden. As a result, these models often do not provide enough insight to the classification, which would be needed to employ them in sensitive domains.

The demand for explainable or interpretable models has been especially pronounced in the medical domain [18, 36, 38, 40]. For example, it is not only of great significance to predict the clinical outcome of a patient, but also to take features of this patient (e.g., age, drug use) into account in an explainable and quantifiable manner [26]. Moreover, models should ideally provide actionable advice for prevention [43]. Simply classifying a patient into a certain health status is not very helpful. The explanations of what has to be improved to change an undesirable status or the identification of early risks determine the usefulness of a model [39].

The most common explanations for classification models are feature importances [3]. Similar to [10], we use the term *feature importance* to describe how important

✉ Mirka Saarela, mirka.saarela@jyu.fi | [1]Faculty of Information Technology, University of Jyvaskyla, University of Jyvaskyla, P.O. Box 35, 40014 Jyväskylä, Finland.

Published online: 03 February 2021

the feature was for the classification performance of the model. More precisely, we refer to feature importance as a measure of the individual contribution of the corresponding feature for a particular classifier, regardless of the shape (e.g., linear or nonlinear relationship) or direction of the feature effect [10, 15]. This means that the feature importances of the input data depend on the corresponding classification model and that a feature important for one model may be unimportant for another model.

Generally, feature importances can be divided into *modular global* and *local* importances [18, 26]. While a modular global feature importance measures the importance of the feature for the entire model, a local importance measures the contribution of the feature for a specific observation. An example of modular global feature importances are the coefficients in L1 regularized logistic regression. L1 regularized logistic regression assigns coefficients based on the importance of a feature, forcing coefficients of unimportant features to exactly zero and providing a magnitude and direction for the remaining coefficients that directly allow an interpretation of the corresponding features.

A local feature importance, in comparison, refers to the contribution of a feature to the results of a trained model on a specific input. An example of the latter are the local interpretable model-agnostic explanations (LIME) developed by Ribeiro et al. [30]. LIME provides features and rules of features that were important for classifying a specific observation. It accomplishes this by learning locally weighted linear models on the neighborhood data of this specific observation to explain its class in an interpretable way. Thus, local feature importances for two individuals can be very different and both might vary from the modular global feature importances.

The purpose of this paper is to compare different classification explanations (i.e., feature importances) for tabular data from the medical domain. The first medical data set we analyze is the well-known publicly available breast cancer data. The second data set is from the field of sports medicine and includes recently collected running injury data. The comparison of explanations is realized by building a linear (logistic regression with L1 penalization) and a non-linear (random forest) model and utilizing their coefficients (logistic regression) and feature importances (random forest) respectively. In addition, for both models the most interesting cases are explained using LIME. LIME is model-agnostic, that is, it can be applied to any nonlinear or linear classifier.

Our research questions are

- What features are the most important?
- Do the most important features differ depending on the technique and if so, which technique should we trust?
- When can local explanations enhance the global modular explanations of a model and should be reported in medical studies?

The motivation of this study is two-fold. First, we empirically test the classification performance of the linear and non-linear classifier. Second, through using modular global and local feature importance techniques we comprehensively compare the explanations provided by these different classifiers. Thus, we are directly addressing a research need pointed out by Tjoa et al. in their 2020 review paper [36]. According to them, the number of medical studies addressing explainability is limited and more studies should compare existing explainability methods.

The remainder of this paper is organized as follows: Sect. 2 provides a short literature review on explainabiblity in machine learning. Section 3 discusses the medical data we used for our analysis. Section 4 explains our analysis framework as well as all used techniques. Section 5 presents the results. Finally, Sect. 6 answers our research questions and summarizes the main findings and implications of this study.

2 About explainability in machine learning

Our work broadly falls under the new research direction of *explainable artificial intelligence* (XAI). XAI refers to approaches and methods that attempt to explain machine learning decisions and predictions in such a way that human domain experts can understand them. Several XAI review papers were published in recent years [17, 18, 24, 36]. According to the 2020 XAI survey by Tjoa et al. [36], XAI is a *young research field* that emerged along the research progress in machine learning and the need to *justify* the decisions and predictions made by the machine learning techniques, particularly if these techniques are applied in sectors that require a top level of *accountability and transparency*. For example, interpretable models are required more in high stake (such as prison sentencing, medical diagnosis, or loan decisions) than low stake (such as movie recommendations) applications [2].

Different taxonomies for XAI techniques have been introduced. In our article, we focus on feature importance or saliency techniques, that is, techniques that explain the decision of an algorithm by assigning values that reflect the importance of input components in their contribution to that decision [36]. These feature importance techniques can be divided into modular global and local techniques. As explained in the introduction, a modular feature importance attempts to describe the importance of the feature for the entire model, while a local feature importance describes the importance of that feature for a

specific input. Moreover, one distinguishes model-specific and model-agnostic techniques [26]. Feature importance techniques that work only for (classes of) particular models are *model-specific*. Feature importance techniques that can be used for any machine learning model and that are applied after model training, are *model-agnostic*. In this paper, we are comparing the following explanations: feature importances of i) logistic regression (modular global and model-specific), ii) random forest (modular global and model-specific), iii) LIME after logistic regression (local and model-agnostic), and iv) LIME after random forest (local and model-agnostic).

Although related and partially overlapping in their techniques, there is a difference between feature importances and feature selection. *Feature selection* is a preprocessing technique [29]. It refers to the general process of detecting the relevant features and discarding the irrelevant ones [4, 34]. One can distinguish filter, wrapper, and embedded feature selection techniques [35]. Filter techniques select features independent of any classifier. Wrapper models utilize the classifier to evaluate on it and find the optimal features. Embedded techniques search the optimal feature subset during the model building process [4]. The process of penalizing irrelevant features and setting their coefficients to zero is an example of embedded feature selection, and at the same also an example of a modular global model-specific feature importance explaining why some features were not important in a logistic regression model. Thus, feature selection and feature importance sometimes share the same technique but feature selection is mostly applied before or during model training to *select the principal features* of the final input data, while feature importance measures are used during or after training to *explain the learned model*.

3 Data retrieval and preprocessing

We used two different data sets with binary classification tasks. The first data set is the openly available breast cancer data from the UCI Archive.[1] This set includes benign and malignant cell samples from 569 patients, 212 with cancer and 157 with fibrocystic breast masses. Each sample contains thirty features, ten real valued features for each cell nucleus (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension) [42]. The classes in the breast cancer data are linearly separable, making the classification a simple task.

[1] The data can be downloaded from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

The second set includes running injury data collected at the University of Calgary. These data are more complex and the classification into healthy or knee-injured runners is most probably a non-linear problem. Running kinematic data were queried from an existing database [14] and 207 knee-injured (n=117) and healthy (n=90) runners (92F, 40.05±14.48 age years) were included in this study. Three-dimensional motion data were collected using an 8-camera motion capture system (MX3+, Vicon Motion Systems, Oxford, UK) while participants ran on a treadmill (Bertec Corporation, Columbus, OH). Spherical retro-reflective markers (9 mm diameter, Mocap Solutions, Huntington Beach, USA) were placed over anatomical landmarks as described in [28]. Joint angles were extracted using 3D GAIT custom software (Running Injury Clinic Inc., Calgary, Alberta, Canada), and time normalized to 101 datapoints per gait cycle (stance and swing), as described more detailed in [27]. In addition, runners participated in strength and flexibility tests. Altogether, the data set contains 154 features.

The breast cancer data has no missing value but some features in the running injury data were not measured for a large part of the runners. To deal with the missing values, we excluded all features that had more than 5% of values missing, ending up with 85 features for the running injury data. The rest of the missing values seemed to be missing at random (MAR) [25] and were imputed using k-nearest neighbour (knn) imputation. Knn imputation works by finding the k most similar observations and then imputing the missing value with a summary metric from those k observations. We used Euclidean distance to measure similarity of observation and a value $k = 10$, as recommended in [37] and imputed with the mean of observations. Both data sets were normalized so that each column had a mean of zero and standard deviation of one with the formula

$$x' = \frac{x - \bar{x}}{\sigma}, \tag{1}$$

where $\bar{x}$ is the mean and σ the standard deviation of the observations. All preprocessing and data analysis was performed with Python 3.6, using scikit-learn and LIME libraries. For knn-imputation a MATLAB (R2018b) script was called.

4 Methods

4.1 L1 regularized logistic regression

L1 regularized logistic regression works by penalizing the feature coefficients with the L1 norm, shrinking some of the feature coefficients to exactly zero. Consider

datapoints $\{(x_i, y_i), i = 1, \ldots, N\}$, where N is the number of observations in data and $x_i \in \mathbb{R}^d$, d is the number of features in data, and $y_i \in \{0, 1\}$ is a binary class label. For classification, the probability of an observation x belonging to class y is given as $P(y|x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}}$, where β is a vector containing d feature coefficients and β_0 is the intercept term.

The cost function to be minimized can be formulated as the negative of the regularized log-likelihood function:

$$\mathcal{L}(\beta_0, \beta) = -\sum_{i=1}^{N} \Big[y_i log(P(y|x)) + (1 - y_i) log(1 - P(y|x)) \Big] + \lambda \sum_{j=1}^{d} |\beta_j|. \quad (2)$$

The last term in the equation is a regularization parameter that is simply the sum of the L1 norms of the feature coefficients where λ controls the strength of regularization. The greater the value of parameter λ, the more coefficients are shrunk to exactly zero. Having less features included makes the model more simple and interpretable. The magnitude of feature coefficients can be interpreted as the importance of that feature, a larger coefficient meaning the feature had more relevance in the classification. In addition, the direction of the coefficient tells whether the feature increases or decreases the probability of belonging to a certain class. The model was trained with the LogisticRegressionCV function and five-fold cross-validation to choose the amount of penalization to use.

4.2 Random forest

Random forest is a nonlinear classification and regression method that is based on building an ensemble of decision trees [8]. Decision trees are tree-like models, where data is split recursively at each decision node into subsets using some rule. The leaf nodes represent the outcome for the observation. The predicted outcome of a random forest model is the mode or mean of the predictions (majority vote) from the individual trees.

Random forests have become very popular, especially in medicine [6, 12, 33], as despite their nonlinearity, they can be interpreted. They provide feature importance measures by calculating the Gini importance, which in the binary classification can be formulated as [23]

$$Gini = p_1(1 - p_1) + p_2(1 - p_2), \quad (3)$$

where p_1 and p_2 are the probabilities of class 1 and 2. The Gini index is minimized when either of the probabilities approaches zero and a total decrease in Gini index (node impurity) is calculated after each node split and then averaged over all trees. The more impurity decreases, the more important the input feature is. The model was trained with the *RandomForestClassifier* function. The maximum number of features to sample at each node and the minimum number of samples required to be at a leaf node were selected with *GridSearchCV* using five folds and values (3,5,9,11) and (1,5,20), respectively.

4.3 Local interpretable model-agnostic explanations

LIME [30] is a recently developed tool providing local interpretability on top of any supervised algorithm. It works by weighting neighbouring observations by their proximity to the observation being explained. The explanation is obtained by training a local linear model based on the weighted neighbouring observations. More precisely, if f is the prediction (in our case classification) model, x is the specific observation for which the prediction $f(x)$ should be explained, g is an explanation model, and π_x the proximity of the neighborhood around x, LIME minimizes the objective function

$$\xi = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \quad (4)$$

where Ω penalizes the complexity of g. This means that from the family of all possible explanations G, the explanation g is chosen that is closest to the prediction of f, while the model complexity $\Omega(g)$ is kept low.

The explainer was trained with the *LimeTabularExplainer* function. As looking at every individual observation would be impractical, we decided to focus on the four most interesting observations with LIME. These include the observation correctly classified as benign/healthy with highest probability, correctly classified as malignant/injured with highest probability, misclassified as benign/healthy with highest probability, and misclassified as malignant/injured with highest probability. For each observation, LIME outputs a rule and an importance value for each feature separately.

4.4 Performance estimation

To estimate the performance of the classification models, we used five-fold cross-validation. Inside each fold, training data were normalized and then test data normalization was done using coefficients estimated from the training data. The missing values in the running injury data were imputed inside each fold, separately for training and test data. The performance was measured using area under the receiver operating characteristic curve (AUC-ROC) [7,

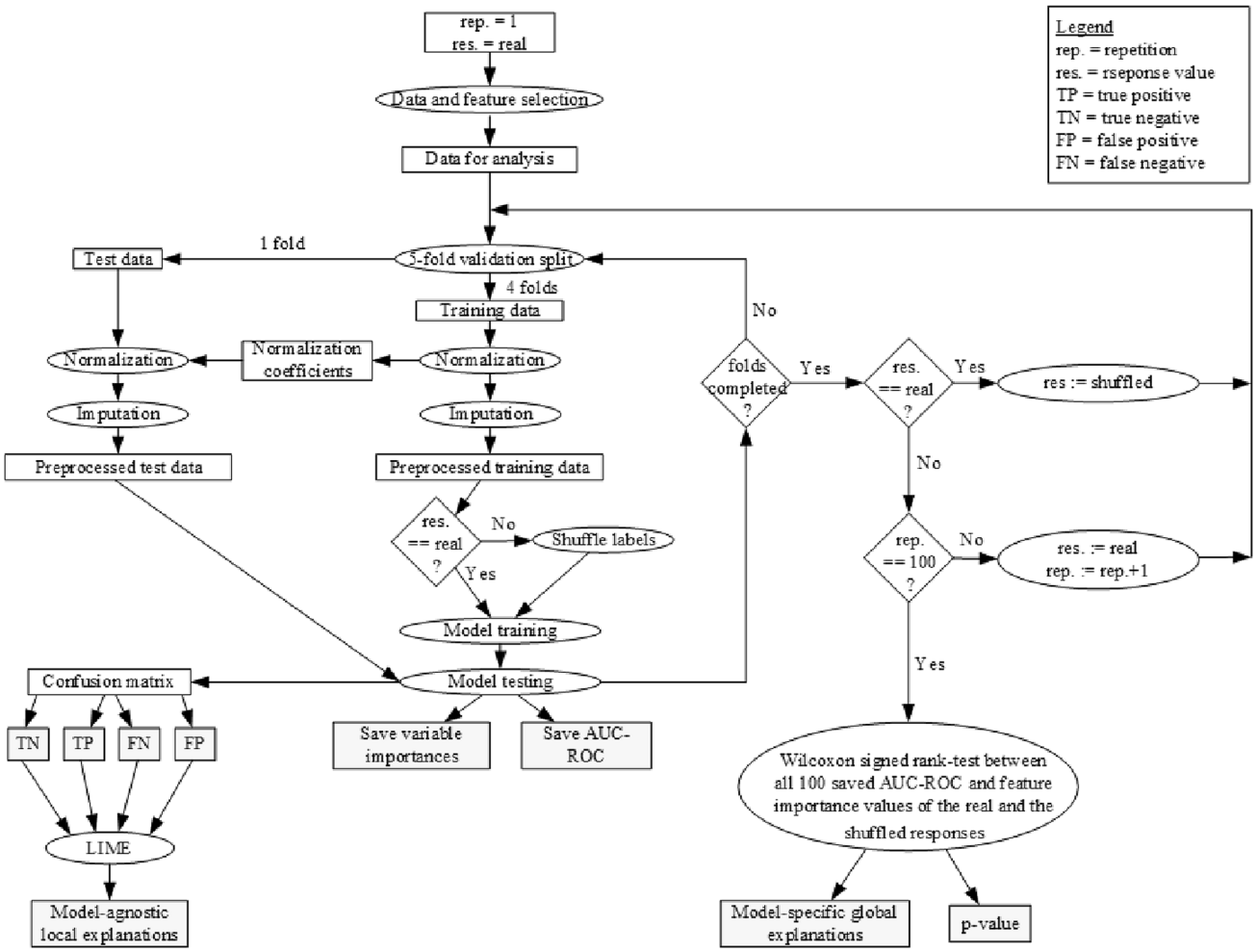

Fig. 1 Flowchart of the applied analysis. For the running injury data, data and feature selection included exclusion of features with more than 5% of missing values as well as inclusion of only knee-injuries from the database for a binary classification task. To identify the most interesting local explanations, the observations classified as true negative, false negative, true positive, and false positive with the highest probability were analyzed with LIME. To identify the significant features of the whole models (modular global explanations), the whole analysis was repeated a hundred times for both the real and the shuffled responses and then, the results were compared with the Wilcoxon signed rank test (cf. [20])

13], averaged over the five folds. Due to random split of folds, results from k-fold validation tend to vary [21, 22]. Therefore, to get a reliable estimate of the performance as well as the feature importances, the whole analysis was repeated a hundred times.

To confirm the significance of the important features and achieved performance, we apply an approach introduced in [20] based on permutations tests. By shuffling the class labels in the training data we made sure that the model was not simply learning some noise in data and therefore achieving higher performance and feature importance values than the chance level [11]. Cross validation splits were the same as in the runs with true labels. Pairwise comparisons of the hundred runs with true and shuffled labels were done with Wilcoxon signed-rank test for the achieved AUC values as well as for the feature importance values of logistic regression and random forest. Limit of significance was set to $\alpha = 0.05$ and Bonferroni corrected. The whole analysis process is outlined in Fig. 1.

5 Results

5.1 Breast cancer data

The feature coefficient and importance values from the classification methods are listed in Table 1. An example plot of feature importance values can be seen in Fig. 2. With logistic regression, all except one (compactness 3) features were significant (Fig. 3). The mean AUC over

Table 1 Logistic regression feature coefficients and random forest feature importances for the breast cancer data

Feature	Coefficient	Importance	Feature	Coefficient	Importance
Radius 1	− 0.088	**0.034**	Texture 1	0.103	0.016
Perimeter 1	− 0.061	**0.041**	Area 1	− 0.020	**0.041**
Smoothness 1	0.061	0.006	Compactness 1	− 0.702	0.011
Concavity 1	**0.831**	**0.042**	Concave points 1	**1.114**	**0.105**
Symmetry 1	− 0.062	0.004	Fractal dimension 1	0.031	0.004
Radius 2	**2.207**	0.014	Texture 2	− 0.396	0.004
Perimeter 2	0.042	0.013	Area 2	0.729	0.032
Smoothness 2	0.211	0.004	Compactness 2	− 0.464v	0.005
Concavity 2	− 0.133	0.007	Concave points 2	0.322	0.005
Symmetry 2	− 0.204	0.004	Fractal dimension 2	**− 0.827**	0.005
Radius 3	**3.170**	**0.113**	Texture 3	**1.723**	0.019
Perimeter 3	0.482	**0.145**	Area 3	0.747	**0.120**
Smoothness 3	**0.573**	0.013	Compactness 3	− 0.089	0.015
Concavity 3	**0.766**	0.032	Concave points 3	1.185	**0.130**
Symmetry 3	**0.692**	0.010	Fractal dimension 3	0.423	0.006

Bolded are the nine features detected with random forest and nine most important features with regression, ranked based on the *p*-value

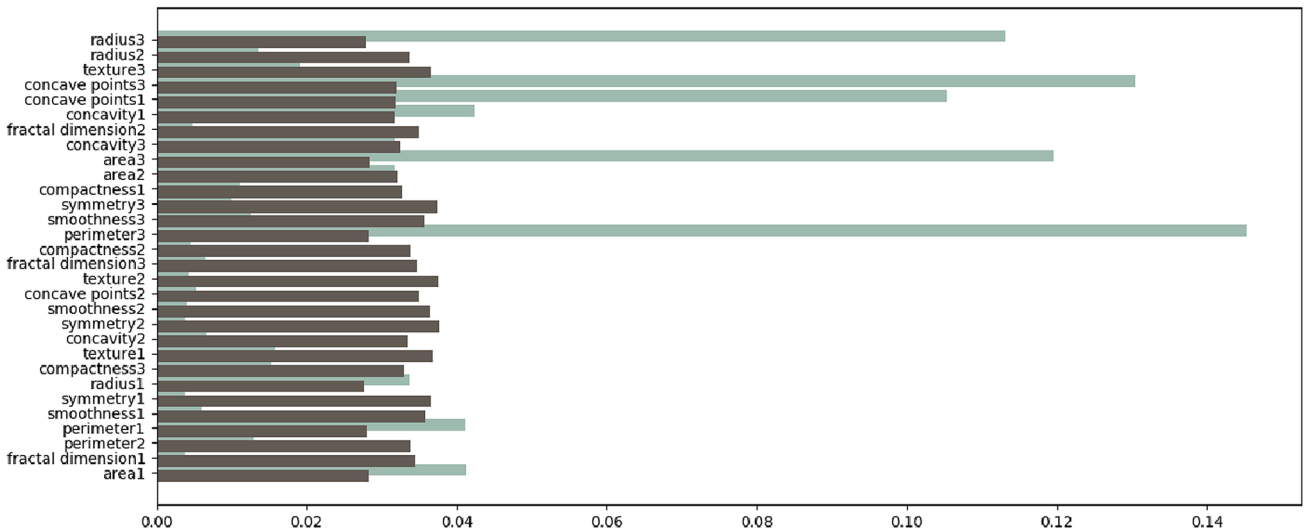

Fig. 2 Modular global feature importance values of the random forest model for breast cancer data. Light bars correspond to the results with real response, darker ones to the results with randomized response. The significant features (i.e., those with a high real feature importance compared to the randomized response) can also be detected visually in this case. One example is the *perimeter 3* feature that was the most significant feature for this model

five folds and hundred repetitions was 0.99 ± 0.002 and for randomized response the AUC values were significantly ($p < 0.001, T = 0.0$) lower (0.50 ± 0.03). The mean AUC for training data was 0.99 ± 0.00. With random forest, nine features were significant and the mean AUC was 0.99 ± 0.001. Again, the randomized runs had a significantly ($p < 0.001, T = 0.0$) lower mean AUC (0.51 ± 0.08). The training AUC was 0.99 ± 0.00.

Nine features (*radius 1*, *perimeter 1*, *area 1*, *concavity 1*, *concave points 1*, *radius 3*, *perimeter 3*, *area 3*, *concave points 3*) were detected by both classification methods. However, if only looking at the set of nine most important features in logistic regression (Table 1), they differ from the set detected by random forest. Feature importance values from LIME for the four assessed observations can be seen in Table 2. For a few observations the set of most important features was largely the same with the classification methods (global explanation) as well as with with LIME (local explanation). However, for most observations new features were detected with LIME.

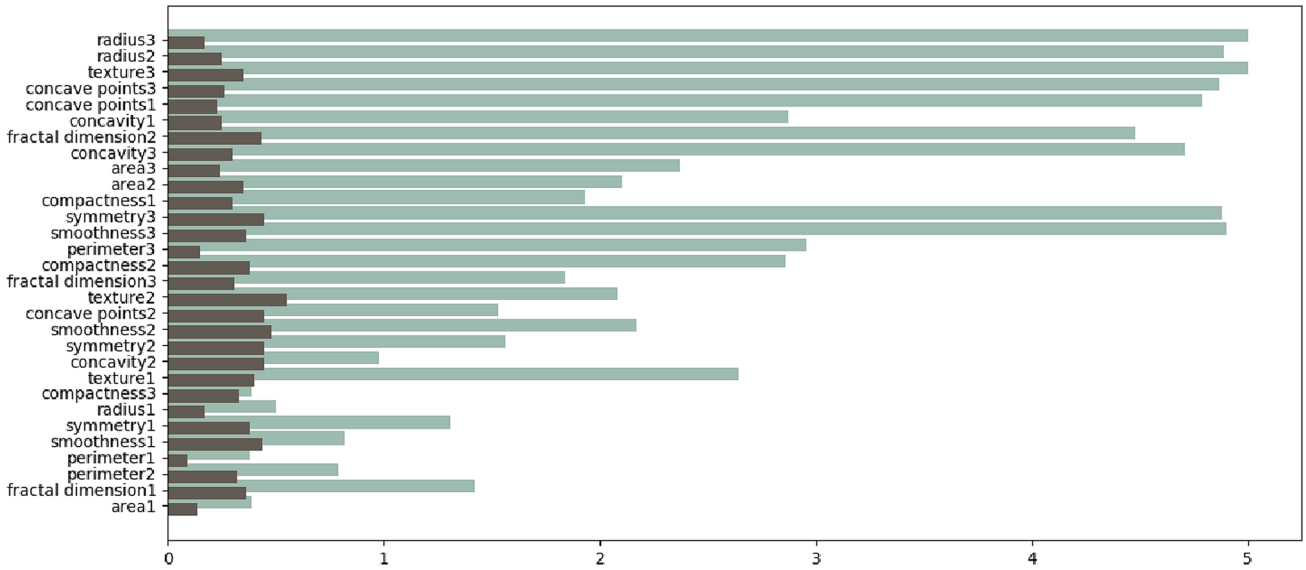

Fig. 3 Modular global feature importance values showing the average of how often (i.e., in each of the 5 folds within the 100 repetitions) a feature was selected by the L1 regularized logistic model for the breast cancer data. Light bars correspond to the results with real response, darker ones to the results with randomized response. Some features, such as the *radius 3*, were chosen almost every time with the real response, indicating that they are extremely important for the model

For all ten correctly classified observations (benign/malignant from five folds) with random forest, the four most important LIME features were also recognized by random forest. In addition, the four most important LIME features were the same for all ten observations, just in different orders. However, for those misclassified there were some observations where the top 4-5 most important LIME features were not detected by random forest. Especially in the case of misclassified as bening, LIME could provide very beneficial information on why that specific observation was not recognized as malignant. On average out of the top 9 LIME features, 5.1 were detected with random forest as well.

In general, as logistic regression detected all but one of the features, the most important LIME features were detected by the method as well. Compared to random forest, the most important LIME features were not as consistent between the ten correctly classified observations; the set of four most important LIME features included 13 different features.

5.2 Running injury data

The feature coefficient and importance values from the classification methods are listed in Table 3. With logistic regression, 61 features were found significant. The mean AUC was 0.70 ± 0.03, while for randomized response the AUC values were significantly ($p < 0.001, T = 0.0$) lower (0.50 ± 0.02). The mean AUC for training data was 0.89 ± 0.03. With random forest, 22 features were detected. The mean AUC was $0.74 \pm 0.01$and again, the randomized runs had a significantly ($p < 0.001, T = 0.0$) lower mean AUC (0.51 ± 0.06). The training AUC was 0.95 ± 0.02.

With the running injury data, features detected with classification methods differed more; out of the 22 features detected by random forest, only 13 were among the 61 detected by logistic regression. Both methods were quite consistent in choosing the features from both legs. The 13 features detected by both classification methods are *age, run level, left hip abductor and right hip external rotation strength, right hip internal rotation flexibility, knee flexion peak of both legs, left knee adduction excursion and pelvis drop peak, left stride rate, swing time and both right and left stance time*.

Feature importance values from LIME for the four assessed observations can be seen in Table 4. With random forest, the four most important LIME features of the ten correctly classified (healthy/injured from five folds) observations were also recognized by the method. For the ten misclassified, all but one also had their four most important LIME features among those detected by random forest. So with this data, the locally and globally most important features seem to be similar. On average out of the top 22 LIME features, 16 were detected with random forest as well.

Again, as logistic regression detected most of the features (61/85), the most important LIME features were detected by the regression method as well. On average out of the top 61 LIME features, 42 were detected with logistic regression as well.

Table 2 LIME results for four people from the first fold on columns, nine most important features from logistic regression (LR) and then random forest (RF) on rows

Correctly classified bening LR		Correctly classified malignant LR	
Radius2	0.23	**Concavity1**	0.34
Fractal dimension2	− 0.21	Area2	0.25
Concave points2	0.19	Radius2	0.23
Fractal dimension3	0.18	**Concave points1**	0.18
Compactness1	− 0.17	Concave points2	0.17
Texture3	− 0.15	**Radius3**	0.15
Radius3	− 0.15	Texture3	0.15
Symmetry2	− 0.14	Fractal dimension2	0.13
Compactness2	− 0.13	**Area3**	0.12
Misclassified bening LR		Misclassified malignant LR	
Concavity1	0.33	Texture3	0.16
Fractal dimension2	− 0.24	**Symmetry3**	− 0.11
Compactness1	− 0.18	Symmetry2	0.08
Concave points2	0.18	Fractal dimension3	0.07
Fractal dimension3	0.18	Symmetry1	0.06
Symmetry2	− 0.15	Concave points2	− 0.04
Symmetry3	0.14	Texture2	-0.04
Compactness2	− 0.14	**Concave points3**	− 0.03
Texture3	0.14	Compactness1	− 0.03
Correctly classified bening RF		Correctly classified malignant RF	
Area3	− 0.08	**Perimeter3**	0.14
Perimeter3	− 0.08	**Area3**	0.13
Radius3	− 0.07	**Concave points3**	0.13
Concave points3	− 0.07	**Radius3**	0.12
Texture3	− 0.05	Area2	0.07
Concave points1	− 0.03	**Concave points1**	0.06
Concavity3	− 0.03	Texture3	0.05
Area2	− 0.03	**Area1**	0.05
Texture1	− 0.02	Texture1	0.04
Misclassified bening RF		Misclassified malignant RF	
Perimeter3	− 0.08	**Area3**	0.14
Area3	− 0.08	**Perimeter3**	0.14
Radius3	− 0.06	**Radius3**	0.12
Texture3	0.06	**Concave points3**	− 0.07
Concavity3	0.04	area2	0.06
Area2	− 0.04	**Area1**	0.05
Smoothness3	0.02	Texture3	− 0.05
Area1	− 0.01	Texture1	− 0.04
Concave points1	0.01	Concavity3	− 0.03

Bolded are those features that were detected also by both classification methods

6 Discussion

The purpose of this paper was to compare explanation measures for linear and non-linear classification models in the medical field. Similarly to previous studies, we also found that the non-linear method (random forest) outperformed the linear method (L1 penalized logistic regression). However, despite the general notion that linear

Table 3 Logistic regression feature coefficients and random forest feature importances for the running injury data

Feature	Coefficient		Importance	
	Left	Right	Left	Right
Q angle	0.170	**– 0.580**	0.005	0.004
Leg length	– 0.086	0.042	0.006	0.006
Hip abduction strength	**– 0.894**	0.095	**0.058**	**0.050**
Hip internal rotation strength	– 0.067	– 0.064	**0.024**	**0.017**
Hip external rotation strength	**0.498**	-0.355	0.005	0.006
Hip internal rotation flexibility	– 0.046	**0.646**	**0.012**	**0.023**
Hip flexion flexibility	0.247	0.025	0.008	0.005
Hip external rotation flexibility	0.163	– 0.172	**0.015**	0.009
IT band flexibility	**– 0.306**	– 0.129	0.008	0.008
Dorsiflexion peak	– 0.065	– 0.160	0.007	0.009
Ankle eversion peak	0.116	– 0.113	0.010	0.006
Ankle eversion pct of stance	– 0.174	**0.851**	0.005	0.005
Ankle eversion excursion	**0.352**	– 0.330	0.011	0.006
Ankle eversion velocity peak	0.093	0.154	0.010	0.011
Ankle pronation onset	0.622	**– 0.737**	0.005	0.004
Ankle pronation offset	– 0.073	-0.031	0.005	0.006
Ankle progression angle	– 0.072	0.043	0.007	0.011
Foot heelstrike angle	0.176	– 0.034	0.009	0.012
Hip extend peak	– 0.048	0.211	**0.019**	0.010
Hip adduction peak	0.194	**– 0.294**	0.008	0.008
Hip adduction excursion	– 0.109	0.188	0.009	**0.021**
Hip abduction velocity peak	**0.331**	**-0.432**	0.010	0.007
Hip adduction velocity peak	– 0.035	– 0.044	0.006	**0.014**
Knee flexion peak	**– 0.209**	**– 0.516**	**0.022**	**0.030**
Knee adduction peak	0.123	**0.348**	0.006	0.008
Knee adduction excursion	– 0.286	0.115	0.007	0.008
Knee adduction velocity peak	0.021	0.098	0.005	0.006
Knee abduction peak	0.011	– 0.209	0.005	0.008
Knee abduction excursion	0.032	– 0.150	0.005	0.005
Knee abduction velocity peak	0.114	**0.385**	0.009	0.012
Pelvis drop peak	**0.588**	-0.174	**0.024**	0.006
Pelvis drop excursion	0.269	0.343	0.009	0.006
Pelvis drop velocity peak	0.027	**– 0.397**	0.006	0.007
Step width	**– 0.093**	**– 0.093**	0.007	0.007
Stride rate	– 0.160	– 0.048	**0.025**	**0.025**
Stride length	– 0.094	– 0.109	0.007	0.007
Swing time	**0.433**	– 0.402	**0.011**	0.007
Stance time	– 0.245	0.232	**0.017**	**0.018**
Heel whip excursion toe off	**0.284**	0.021	0.008	0.007
Vertical oscillation	– 0.008	0.089	**0.025**	**0.017**
Height cm	0.080		0.004	
Sex	0.221		0.002	
Weight	– 0.108		0.006	
Age	**– 0.751**		**0.065**	
Run level	– 0.070		**0.026**	

Bolded are the 22 features detected with random forest and 22 most important features with regression, ranked based on the *p*-value. *Left* are the features measured from left leg; *Right* from right leg. The last five rows contain general demographic features

models provide better interpretability, we also found that the non-linear method offered a better explainability as it selected fewer features in the analyzed cases.

The logistic regression feature importances were harder to interpret. More penalization would result less features in the model but then the performance might decrease even more. Moreover, if there are highly correlated features logistic regression might just arbitrary choose one of those [5]. Another point to consider when comparing the techniques is that feature coefficients in logistic regression are calculated with all features as input in the model, while random forest calculates the importance values separately for each feature.

To answer our first research question: The feature importance measures selected different features but overall, *radius, perimeter, area*, and *concave points 3* and *1* as well as *concavity 1* were the features selected by both for breast cancer data while for the running injury data, most important features included knee and hip/pelvis features as well as age, running level and some functional features. A previous study found 14 important features in the breast cancer data with a genetic algorithm [1], out of which only three were among the nine most important in this study. This shows that different features are detected with different methods. In addition, the features are highly correlated and the classification task very simple in the breast cancer data, so high accuracy can be achieved using different sets of features.

This also answers the first part of our second research question. The most important features indeed differed depending on the used technique. Concerning the second part of our second question, our experiments seem to provide better results for random forest. However, we believe that a triangulated approach [31] of several techniques would enhance trust the most. As already emphasized by Gifi [16] if different techniques lead to the same conclusion, it is more likely that these reflect genuine aspects of the data.

With regard to our last question we think that in the medical domain, those cases that were classified wrong as benign/healthy (false negatives) are of most interest and should be accompanied with local explanations. False positives are also interesting, but a little less critical than false negatives. The most important features for misclassified cases are of interest as they are misleading.

Naturally, the results are limited to the used data and techniques. In future work, we intend to repeat the presented analysis scheme in a larger scale. More specifically, we are interested in comparing the techniques utilized here to such as neural networks. Using more complex non-linear models can lead to better performance with large real world data sets while LIME can provide interpretability for the results.

Table 4 LIME values, four people from first fold in separate cells and nine most important features from logistic regression and random forest on rows in each cell

Correctly classified healthy LR		Correctly classified injured LR	
df peak L	− 0.24	df peak L	0.22
Knee add excur L	0.10	eve vel peak R	0.15
q angle L	− 0.09	Drop vel peak L	0.12
Hip ext rot F R	− 0.07	Heel whip excur L	0.11
Hip add vel peak R	− 0.06	eve pct stance L	0.10
Drop vel peak L	− 0.06	**knee add excur L**	0.08
eve vel peak R	0.05	Hip add vel peak R	− 0.06
Stance time L	0.04	Hip add excur R	− 0.07
Hip int rot S L	0.04	Pron offset L	0.05
itband R	− 0.04	**Vert osc R**	− 0.04
Pron onset R	− 0.03	Pron onset R	− 0.04
Heel whip excur L	− 0.02	itband R	0.03
Misclassified healthy LR		Misclassified injured LR	
eve vel peak R	− 0.14	itband R	0.14
Drop vel peak L	− 0.12	Drop vel peak R	− 0.11
eve pct stance L	0.10	Heel whip excur L	0.10
q angle L	0.09	eve pct stance L	0.10
knee add excur L	− 0.07	q angle	0.08
Pron offset L	0.07	**Hip int rot F R**	0.08
Hip ext rot F R	− 0.06	**Knee add excur L**	− 0.07
df peak L	− 0.06	Pron offset L	0.06
Hip add vel peak R	− 0.05	df peak L	0.06
Hip add excur R	− 0.06	Knee abd vel peak R	0.05
Step width L	− 0.05	**Stance time L**	0.04
Vert osc R	− 0.04	Hip add excur R	− 0.04
Stance time L	− 0.04	Step width R	− 0.04
Correctly classified healthy RF		Correctly classified injured RF	
Hip abd L	− 0.01	df peak R	0.01
Hip abd R	− 0.01	Drop excur L	0.01
df peak R	− 0.01	df peak L	0.01
Heel whip excur L	0.01	**Flex peak L**	0.01
Drop excur L	− 0.01	**Knee add excur L**	0.01
Drop vel peak R	− 0.01	**Hip abd R**	0.01
knee add excur L	− 0.01	Extend peak R	0.01
drop vel peak R	− 0.01	Heel whip excur L	0.01
Extend peak L	− 0.01	**Extend peak L**	0.01
Extend peak R	− 0.01	itband R	0.01
Heelstrike ang L	− 0.01	**Hip abd L**	0.01
Flex peak L	− 0.01	itband L	0.01
Hip ext rot S R	0.01	step width R	0.01
eve excur L	− 0.01	Leg length L	0.01
Misclassified healthy RF		Misclassified injured RF	
itband R	− 0.02	df peak R	− 0.01
hip abd L	− 0.01	**Hip abd L**	0.01
df peak R	− 0.01	Hip flex R	0.01
itband L	− 0.01	**Sub age**	− 0.01
Flex peak L	− 0.01	**Hip abd R**	0.01
Hip abd R	− 0.01	**Extend peak L**	0.01
Heel whip excur L	0.01	**Flex peak L**	0.01
Drop vel peak R	− 0.01	**Knee add excur L**	− 0.01
Extend peak R	− 0.01	eve vel peak R	− 0.01
Step width R	0.01	itband R	0.01
Drop excur L	− 0.01	**Run level**	0.01
Hip flex R	0.01	Heel whip excur L	− 0.01
Knee abd vel peak R	− 0.01	Leg length L	0.01

Bolded are those that were detected by both classification methods as well. (L=left, R=right, S=strength, F=flexibility)

Acknowledgements This research was supported by the Academy of Finland (Grant No. 311877) and is related to the thematic research area DEMO (Decision Analytics Utilizing Causal Models and Multiobjective Optimization, jyu.fi/demo) of the University of Jyväskylä, Finland.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical approval The running injury data collection was approved by the University of Calgary's Conjoint Health Research Ethics Board (CHREB: REB15- 0557). Before data collection, all participants provided a written informed consent to participate. The breast cancer dataset includes de-identified data from a public repository [45]. As such, ethical approval was not required.

References

1. Aličković E, Subasi A (2017) Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput Appl 28(4):753–763
2. Ashoori M, Weisz JD (2019) In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. arXiv preprint arXiv:1912.02675
3. Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, Puri R, Moura JM, Eckersley P (2020) Explainable machine learning in deployment. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp 648–657
4. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inf Syst 34(3):483–519
5. Bondell HD, Reich BJ (2008) Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64(1):115–123
6. Boulesteix AL, Janitza S, Kruppa J, König IR (2012) Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discover 2(6):493–507
7. Bradley AP (1997) The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recogn 30(7):1145–1159
8. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
9. Breiman L (2001) Statistical modeling: The two cultures. Stat Sci 16(3):199–231
10. Casalicchio G, Molnar C, Bischl B (2019) Visualizing the Feature Importance for Black Box Models. Lect Notes Comput Sci 11051:655–670
11. Combrisson E, Jerbi K (2015) Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 250:126–136
12. Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinf 7(1):3
13. Fawcett T (2006) An introduction to roc analysis. Pattern Recogn Lett 27(8):861–874
14. Ferber R, Osis ST, Hicks JL, Delp SL (2016) Gait biomechanics in the era of data science. J Biomech 49(16):3759–3761
15. Fisher A, Rudin C, Dominici F (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res 20(177):1–81
16. Gifi A (1990) Nonlinear multivariate analysis. Wiley, Hoboken
17. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics, pp 80–89. IEEE
18. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput surv CSUR 51(5):1–42
19. Horn F, Pack R, Rieger M (2020) The autofeat python library for automated feature engineering and selection. In: Cellier P, Driessens K (eds) Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, Cham, pp 111–120
20. Jauhiainen S, Kauppi JP, Leppänen M, Pasanen K, Parkkari J, Vasankari T, Kannus P, Äyrämö S (2020) New machine learning approach for detection of injury risk factors in young team sport athletes. International journal of sports medicine
21. Kohavi R, et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, vol. 14, pp. 1137–1145. Montreal, Canada
22. Krstajic D, Buturovic LJ, Leahy DE, Thomas S (2014) Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform 6(1):10
23. Kuhn M, Johnson K et al (2013) Applied predictive modeling, vol 26. Springer, Berlin
24. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever hans predictors and assessing what machines really learn. Nat Commun 10(1):1–8
25. Little RJ, Rubin DB (2014) Statistical analysis with missing data, vol 793. Wiley, Hoboken
26. Molnar C (2019) Interpretable Machine Learning. Lean Publishing
27. Phinyomark A, Hettinga BA, Osis ST, Ferber R (2014) Gender and age-related differences in bilateral lower extremity mechanics during treadmill running. PLoS ONE 9(8):e105246
28. Pohl MB, Lloyd C, Ferber R (2010) Can the reliability of three-dimensional running kinematics be improved using functional joint methodology? Gait Posture 32(4):559–563
29. Remeseiro B, Bolon-Canedo V (2019) A review of feature selection methods in medical applications. Comput Biol Med 112:103375
30. Ribeiro MT, Singh S, Guestrin C (2016) "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, pp 1135–1144
31. Saarela M (2017) Automatic knowledge discovery from sparse and large-scale educational data: case Finland. 262. University of Jyväskylä

32. Saarela M, Kärkkäinen T (2020) Can we automate expert-based journal rankings? Analysis of the Finnish publication indicator. J Inf 14(2):101008
33. Saarela M, Ryynänen OP, Äyrämö S (2019) Predicting hospital associated disability from imbalanced data using supervised learning. Artif Intell Med 95:88–95
34. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517
35. Tang J, Alelyani S, Liu H (2014) Feature selection for classification: A review. Data classification: Algorithms and applications p 37
36. Tjoa E, Guan C (2020) A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst pp 1–21
37. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for dna microarrays. Bioinformatics 17(6):520–525
38. Vellido A (2019) The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl pp 1–15
39. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL Tech. 31:841
40. Waring J, Lindvall C, Umeton R (2020) Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artif Intell Med 104:101822
41. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12(4):e0174944
42. Wolberg WH, Street WN, Mangasarian O (1994) Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Lett 77(2–3):163–171
43. Yang C, Delcher C, Shenkman E, Ranka S (2016) Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In: 2016 IEEE 18th International conference on e-Health networking, applications and services (Healthcom), pp 1–6. IEEE
44. Zien A, Krämer N, Sonnenburg S, Rätsch G (2009) The feature importance ranking measure. Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin, pp 694–709
45. Zwitter M, Soklic M (1988) UCI machine learning repository breast cancer wisconsin data. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer