Received: 7 April 2020 | Revised: 8 October 2020 | Accepted: 14 October 2020

DOI: 10.1002/ejp.1683

ORIGINAL ARTICLE

EJP European Journal of Pain

Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (XAI)

Jörn Lötsch[1,2] | **Sebastian Malkusch**[1]

[1]Institute of Clinical Pharmacology, Goethe - University, Frankfurt am Main, Germany

[2]Fraunhofer Institute for Translational Medicine and Pharmacology ITMP, Frankfurt am Main, Germany

Correspondence
Jörn Lötsch, Goethe – University, Theodor Stern Kai 7, 60590 Frankfurt am Main, Germany.
Email: j.loetsch@em.uni-frankfurt.de

Funding information
This work has been funded by the Landesoffensive zur Entwicklung wissenschaftlich-ökonomischer Exzellenz (LOEWE), LOEWE-Zentrum für Translationale Medizin und Pharmakologie (JL), in particular through the project "Reproducible cleaning of biomedical laboratory data using methods of visualization, error correction and transformation implemented as interactive R-notebooks " (JL).

Abstract

Background: In pain research and clinics, it is common practice to subgroup subjects according to shared pain characteristics. This is often achieved by computer-aided clustering. In response to a recent EU recommendation that computer-aided decision making should be transparent, we propose an approach that uses machine learning to provide (1) an understandable interpretation of a cluster structure to (2) enable a transparent decision process about why a person concerned is placed in a particular cluster.

Methods: Comprehensibility was achieved by transforming the interpretation problem into a classification problem: A sub-symbolic algorithm was used to estimate the importance of each pain measure for cluster assignment, followed by an item categorization technique to select the relevant variables. Subsequently, a symbolic algorithm as explainable artificial intelligence (XAI) provided understandable rules of cluster assignment. The approach was tested using 100-fold cross-validation.

Results: The importance of the variables of the data set (6 pain-related characteristics of 82 healthy subjects) changed with the clustering scenarios. The highest median accuracy was achieved by sub-symbolic classifiers. A generalized post-hoc interpretation of clustering strategies of the model led to a loss of median accuracy. XAI models were able to interpret the cluster structure almost as correctly, but with a slight loss of accuracy.

Conclusions: Assessing the variables importance in clustering is important for understanding any cluster structure. XAI models are able to provide a human-understandable interpretation of the cluster structure. Model selection must be adapted individually to the clustering problem. The advantage of comprehensibility comes at an expense of accuracy.

1 | INTRODUCTION

Computer-aided decision making has become common in biomedical research and clinical practice, including the implementation of artificial intelligence (AI) techniques, most often as machine-learning algorithms referred to as a set of methods that can automatically detect patterns in data and then use the uncovered patterns to predict or classify future data, to observe structures such as subgroups in the data or to extract information from the data suitable to derive new knowledge

(Dhar, 2013; Murphy, 2012). Machine-learning algorithms appear in medicine in two different flavors: sub-symbolic, which function like a black box, and symbolic, which provide comprehensible decision-making processes. Recent advances in machine learning have triggered the publication by the European Union (EU) of landmark papers emphasizing the need for computer-based decisions to be transparent so that they can be communicated in an understandable way to the patients they affect (Hamon et al., 2020). To address this problem, the concept of explainable AI (XAI) is attracting increasing scientific interest (Arrieta et al., 2019).

Among computer-assisted methods in pain research, the detection of subgroups in patients has long been used. Given the current efforts towards precision medicine in pain, it is expected that treatments tailored to the individual needs of patients will provide clinical guidance with positive effects on outcomes (Fillingim et al., 2014) based on knowledge of various factors relevant to the individual development of pain and its impact on life (Kaiser et al., 2018; Vartiainen et al., 2016), such as comorbidities, neurobiological or psychological factors. This requires an assessment of whether the patients represent a heterogeneous cohort, and by which features possible subgroups are characterized. For this purpose, methods of data projection (Lötsch & Ultsch, 2019) and cluster identification (Ultsch & Lötsch, 2017) are available. Clustering is widely used in pain research. A PubMed database search on https://pubmed.ncbi.nlm.nih.gov for "(pain AND (clustering OR "cluster analysis")) NOT (review[Publication Type]) NOT ("cluster trial" OR "cluster randomized trial" OR "cluster randomized controlled trial" OR "cluster headache")" produced 4,954 results on 8 October 2020.

Variable importance is a key to the understanding of a cluster structure. In order to take into account the EU advice mentioned above, there is a need for (1) a comprehensible interpretation of a cluster structure up to (2) a transparent decision process that can be communicated to the person concerned explaining why the particular person is placed in a particular cluster. Classical approaches such as the interpretation of the results of a principal component analysis (PCA) can only partially achieve this. Therefore, the present report proposes a machine-learning–based approach that uses XAI methods (Ribeiro et al., 2016) for the interpretation of pain-related cluster structures that may have been detected by various methods, some of which have been discussed previously in a pain context (Ultsch & Lötsch, 2017).

Significance

Explainable artificial intelligence is a suitable method for the interpreation of clusters or subgroups emerging from complex hogh-dimensional pain-related data.

2 | METHODS

2.1 | Data sets

A suitable pain-related data set was available from a recent study on patterns in parameters of quantitative sensory testing (QST) in response to harmful thermal stimuli (Lötsch et al., 2017, 2020). However, its cluster structure was initially unknown and had to be determined using a clustering method. In order to evaluate the correct functioning of cluster interpretation methods, the true cluster structure should be known, which is difficult to obtain with pain-related data sets. Therefore, two other data sets providing such known cluster or subgroup structures were included; both have been widely used for pattern recognition problems in any context, namely, the iris flower data set (Fisher, 1936) and the "Fundamental Clustering and Projection Suite" (FCPS (Ultsch & Lötsch, 2020)).

2.1.1 | Pain thresholds to thermal stimuli

Pain thresholds to thermal stimuli were acquired in a cohort that originally consisted of $n = 100$ healthy volunteers (46 men) of Caucasian ethnicity, after self-assignment, aged between 19 and 42 years (mean ± standard deviation 25 ± 3.5 years). All volunteers were pain free when the experiments started. Due to missing data and deviations from the original study protocol, the pain-related data of $n = 82$ volunteers were complete. The study followed the Declaration of Helsinki on biomedical research in humans and was approved by the Ethics Committee of the Medical Faculty of the Goethe University Frankfurt am Main, Germany. All test persons had given written consent to the study procedures including genotyping. The details of the study have already been reported in detail (Lötsch et al., 2017, 2020).

As described previously (Lötsch et al., 2017), the recording of sensory thresholds for different stimuli was performed in strict compliance with a standard procedure developed by the German Research Network for Neuropathic Pain (Rolke, et al., 2006; Rolke, et al., 2006). Pain thresholds for harmful heat and cold were selected for the present analyses. They were assessed with a 3 × 3 cm^2 thermode (TSA 2001 - II, Ramat Yishai, Israel) placed on a 9 cm^2 skin area on the inside of the forearm without superficial veins or birthmarks. The heat pain thresholds (HPT) were measured by increasing the temperature of the thermode by 1°C/s, starting at 32°C, until the subject indicated pain, which triggered the reversal of the temperature ramp back to baseline. HPT was defined as the mean value of three repeated measurements. Cold pain

thresholds (CPT) were recorded analogously, except that the temperature of the thermode was lowered from 32°C at −1°C/s to a shut-off temperature of 0°C. In addition, a hypersensitization procedure was applied consisting of ultraviolet (UV-B) light applied to the inner side (Gustorff et al., 2004; Harrison et al., 2004; Hoffmann & Schmelz, 1999). After calibration of the UV-B intensity for the individual skin type by determining the minimum erythema dose (MED), a cumulative UV-B dose between 200 mW/cm^2 and 600 mW/cm^2 (UV-B lamp UV 109 from Waldmann Medizintechnik, Villingen-Schwenningen, Germany) corresponding to 2 MED was administered to the inner side of the forearm 24 h before the actual experiments.

The acquired pain thresholds were z-transformed into a reference group, comprising 180 healthy subjects (Rolke, et al., 2006), separately for test area, gender and age of the subjects. Specifically, the acquired parameter values were z-transformed as $z_{QST,individual} = \frac{QST_{individual} - QST_{reference}}{Standarddeviation_{reference}}$, where the QST mean reference values and standard deviations were the published values (Magerl et al., 2010), with respect to gender, age and tested body part (Appendix 2 in (Rolke, et al., 2006)) of the actual subject. This is possible for individual subjects, i.e. mean values and standard deviations are taken from the reference groups according to the established protocol of the current QST system. It, therefore, allows to analyze the occurrence of pathological values of an individual patient when used in clinical settings for diagnostic purposes of neuropathic pain. The signs of the z-scores were adjusted according to the standard instructions in such a way that a z-score > 0 indicates high sensitivity and a z-score < 0 indicates low sensitivity. These z-scores served as a basis for further analyses.

The z-transformed QST parameters were checked for the tolerable amount of abnormal values (|x_z |≥ 1.96) in a routine quality control based on a binomial distribution (Dimova et al., 2016) according to the standard procedure defined by the German Research Network for Neuropathic Pain for the present standardized QST test system. For a set of 82 samples, the procedure tolerates 18 abnormal values (mild suspicion of measuring errors) and for a one-sided 5% quantile and 21 abnormal values for a one-sided 1% quantile (strong suspicion of measuring errors) (Vollert et al., 2015). Both of the above-mentioned quality assurance requirements were met by the presented QST parameters.

The thresholds for heat stimuli were multiplied by a value of −1 to obtain a uniform direction with larger values indicating high pain sensitivity. The effect of UV-B on the thresholds was quantified as the difference between the measurement after UV-B application and the measurement without UV-B application, i.e. $UVBEff_{Heat} = zHPT_{UVB} - zHPT_{baseline}$ for heat pain and $UVBEff_{Cold} = zCPT_{UVB} - zCPT_{baseline}$ for cold pain.

2.1.2 | Standard non-pain data sets with known cluster structure

In order to test the correct functioning of the present approach to cluster interpretation, standard data sets not related to pain but often used for similar methodological testing purposes were included. The iris flower data set (Anderson, 1935; Fisher, 1936) has often been used for pattern recognition problems. It gives the measurements in centimetres of the variables sepal length and width or petal length and width for 50 flowers each of the three species *iris setosa*, *versicolor* and *virginica*. As there are apparently at least half a dozen different versions of this data set (Bezdek et al., 1999), it is necessary to specify that in the present analysis, the version implemented in R software package (R Development Core Team, 2008) as "data(iris)" was used. The "Fundamental Clustering and Projection Suite" (FCPS (Ultsch & Lötsch, 2020)) provides 10 data sets with one to seven classes with various degrees of separability. One data set, "Golfball", has only one class and was therefore omitted. The FCSP data set collection is freely available at http://www.mdpi.com/2306-5729/5/1/13/s1.

2.2 | Data analysis

Data were analysed using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008) concomitantly on three computers running Ubuntu Linux 20.04 LTS 64-bit (Canonical, London, UK)). The data analysis focused on the evaluation of an approach to understandable cluster interpretation including a transparent process of cluster assignment of a given subject.

2.2.1 | Quantitative variables

Pain-related data comprised z-transformed pain thresholds to heat or cold stimuli recorded under control conditions, after UV-B irradiation and the UV-B effects calculated as the difference between the z-transformed thresholds before and after the irradiation. This provided an 82 × 6 (n × d) sized input data space $D = \{(x_i) | x_i \in X, i = 1, \ldots, n\}$, which included vectors $x_i = <x_{i,1},\ldots,x_{i,d}>$ with $d = 6$ different parameters representing the six different pain thresholds or UV-B effect-related variables ($zHPT_{baseline}$, $zHPT_{UVB}$, $zCPT_{baseline}$, $zCPT_{UVB}$, $UVBEff_{Heat}$, and $UVBEff_{cold}$) acquired from $n = 82$ subjects.

The iris flower data set includes 150 cases and five variables named Sepal.Length, Sepal.Width, Petal.Length, Petal.Width and Species. To test the feature selection capabilities of the present approach, the input variables were duplicated, and the duplicates were randomly permuted. The expectation

was that the algorithm would not select the permuted features. This resulted in a 150 x 8 input data space with d = 8 different features representing the four measurements of iris flowers, either original or randomly permuted, obtained from $n = 150$ flowers. From the FCPS data set collection, nine data sets were used comprising d = 2,…,7 classes and $n = 400$ to 4,096 instances. The data set has been exhaustively described in a separate paper (Ultsch & Lötsch, 2020).

2.2.2 | Interpretation of clusters of pain-related phenotypes

Cluster interpretation by examination of PCA and k-means clustering results

A common clustering approach in pain research includes the creation of uncorrelated variables by orthogonal projection onto a low-dimensional linear space named the principal subspace by means of PCA (Hotelling, 1933; Pearson, 1901), followed by a clustering procedure applied to the cases' projection on the relevant principal components (PCs). Hence, PCA was performed with the d = 6 centred, normalized pain-related variables, using the R-library "FactoMineR" (https://cran.r-project.org/package=FactoMineR (Le et al., 2008)). Of the resulting main PCs, those with eigenvalues >1 were retained for clustering (Guttman, 1954; Kaiser, 1958), which was subsequently performed using the k-means method (MacQueen, 1967) and the Euclidean distance. The number of clusters was obtained as the majority vote among 30 different indices for determining the number of clusters, calculated using the R library "NbClust" (https://cran.r-project.org/package=NbClust). Cluster quality was evaluated by calculating the Silhouette index (Rousseeuw, 1987). The clustering was performed using the R-libraries "flexclust" (https://cran.r-project.org/package=flexclust (Leisch, 2006)) and "cluster" (https://cran.r-project.org/package=cluster (Maechler et al., 2017)).

For cluster interpretation, the contribution of the relevant PCs to the overall variance of the data set and the contributions of each pain-related feature to the PC were calculated by summing up the respective factor loadings. Subsequently, the contribution of PCs to the clusters was addressed directly using an algorithm built to find the variables that control cluster allocation and to evaluate them according to their relevance. It is based on the permutation misclassification rate for each variable, using the mean misclassification rate over all iterations as a measure of the importance of the variables. The algorithm is implemented as "FeatureImpCluster" function in the R library of the same name (https://cran.r-project.org/package=FeatureImpCluster (Pfaffel, 2020)). Cluster interpretation also included standard statistics comprising *t*-tests (Student, 1908) or univariate analyses of variance after non-significant Kolmogorov-Smirnov tests (Smirnov, 1948) against normal distribution; the α level was set at 0.05 and corrected for multiple testing as proposed by Bonferroni (Bonferroni, 1936).

Machine-learning–based cluster interpretation

Methods of supervised machine learning were used for cluster interpretation by transforming the problem of determining the meaning of the cluster into a classification problem. This means that by analyzing which features are needed to assign a case to the correct cluster, the characteristics relevant to the cluster structures become known. Furthermore, by analyzing the decision process of the successful algorithm, the individual cluster assignment becomes transparent. The more successful the correct cluster assignment by the algorithm is, the more valid is the interpretation of the entire cluster substructure of the data set. For this purpose, cluster-relevant pain variables were identified by analyzing on which characteristics the successful class assignment was based by using sub-symbolic classifiers (Smolensky, 2010). Sub-symbolic classifiers usually achieve the highest classification performance by waiving comprehensibility of the process. Subsequently, symbolic classifiers (Newell & Simon, 1976), which include XAI methods, were used to make the decision process of class assignment understandable as a combination of comprehensible rules. It is noteworthy that comprehensibility may be bought with a reduced classification performance (Arrieta et al., 2019). Therefore, the present study exploited the strengths of both types of machine-learning algorithms, Sub-symbolic classifiers (non-XAI type machine learning) for feature selection and symbolic classifiers (XAI-type machine learning) that provide an understandable cluster interpretation based on the selected features. The optimal XAI is chosen from several types of XAI based on their comparative classification performance.

2.2.3 | Selection of cluster-relevant pain-related features using non-XAI type machine-learning methods

Feature selection was approached using a recently proposed method of random forest classification followed by computed ABC analysis (Lötsch & Ultsch, 2020). The selection of random forests was further suggested by its tree-based structure, which it shares with the majority of the XAI methods included in the present analysis. Specifically, random forest classifiers (Breiman, 2001; Ho, 1995) generate sets of different, uncorrelated and often very simple decision trees. The class assignment is achieved as a majority vote across many well-performing trees. Hyperparameters were set after a grid search to forest sizes of 1,500 trees and $\sqrt{(d)}$ parameters for each tree. The calculations were done using the R libraries "randomForest" (https://cran.r-project.org/package=rando

TABLE 1 Characteristics of tree- and rule-based classification algorithms: hierarchical rules, the algorithm returns hierarchical rules; branches/split, number of branches that result from splitting procedure at a node; metrics, split parameter identification; unbiased, the algorithm's choice of a splitting attribute is unbiased; PWI, pairwise interaction detection; Split type: possible splitting procedures at a node; pruning: The algorithm performs post-pruning

Classifier	References	Hierarchical rules	branches/ split	metrics	Unbiased	PWI	Split type	Pruning
ID3	(Quinlan, 1986)	False	≥2	IG	False	False	U	False
C4.5	(Kim & Loh, 2001; Loh, 2011, 2014; Quinlan, 2014; Salzberg, 1994)	True	≥2	IGR	False	False	U	True
C5.0	(Kuhn & Quinlan, 2018; Loh, 2011) (Quinlan, 2014)	True	≥2	IGR	False	False	U	True
CART	(Kim & Loh, 2001; Loh, 2011, 2014) (Breimann et al., 1993)	True	2	Gini Impurity	False	False	U, L	True
CTREE	(Hothorn et al., 2006; Loh, 2014)	False	2	Permutation test	True	False	-	False
CHAID	(Kass, 1980; Loh, 2011, 2014)	False	≥2	χ^2-test	False	False	U	False
QUEST	(Kim & Loh, 2001; Loh, 2011, 2014; Loh & Shih, 1997)	False	2	F-test, χ^2-test	True	False	U, L	True
GUIDE	(Loh, 2009, 2011, 2014)	False	2	χ^2-test	True	True	U, L	True
CRUISE	(Kim & Loh, 2001; Loh, 2011, 2014)	False	≥2	χ^2-test	True	True	U, L	True
FACT	(Kim & Loh, 2001; Loh, 2014; Loh & Vanichsetakul, 1988)	False	≥2	F-test	True(o), False(n)	False	U, L	False
ONER	(Holte, 1993; Kim & Loh, 2001)	True	≥2	χ^2-test	–	–	–	True
RIPPER	(Cohen, 1995)	True	≥2	IG	–	–	–	True
PART	(Frank & Witten, 1998)	True	≥2	IGR	False	False	U	True

Abbreviations: IG, information gain; IGR, information gain ratio; L, linear splits; CTREE, conditional inference trees; Rpart, classification and regression trees (= CART) ; CHAID, chi-square automatic interaction detection; MARS, multivariate adaptive regression splines; CRUISE, classification trees with unbiased multi-way splits; QUEST, quick unbiased efficient statistical trees; PART, partial decision trees; RIPPER, repeated incremental clipping for error reduction; RF, random forests; n, in the case of nominal variables; o, in the case of ordered variables; PLR, piecewise linear regression; RER, residual error reduction; U, univariate splits.

mForest (Liaw & Wiener, 2002)) or "caret" (https://cran.r-project.org/package=caret (Kuhn, 2018)).

In order to identify the features relevant to cluster assignment, random forests were first trained on 100 data subsets obtained by class-proportional Monte Carlo random resampling of 2/3 of the original data. The trained algorithm was then applied to the remaining 1/3 of the data. The performance of class assignment was quantified using the balanced accuracy (Brodersen et al., 2010) as the main criterion. This was preferred to alternatives such as the area under the receiver operator curve (AUC-ROC (Peterson et al., 1954)) based on the assessments in Brodersen et al. (2010). Moreover, the balanced accuracy as the mean of sensitivity and specificity is half of the Youden index (Youden, 1950), calculated as the sum of the two measures, which is often used as a performance measure in machine learning.

The procedure was then repeated, successively leaving out one of the features from the analysis. For each feature, the decrease in balanced classification accuracy when left out of the analysis was maintained providing a quantitative measure of its importance. Using computed ABC analysis (Ultsch & Lötsch, 2015), in each run, the features were categorized into three non-overlapping subsets called "A", "B" and "C" (Juran, 1975; Pareto, 1909). Subset "A" contains the "important few", subset "C" contains the "trivial many" of negligible information value and subset "B" contains items with balanced profit and loss when selected. In the present analyses, the elements placed in subsets "A" and "B" were retained. The final size of the feature set corresponded to the most common size of subsets "A" + "B" in the 100 runs, and the members of the feature set were selected in descending order of their appearances in the retained ABC sets. These calculations were performed using our R package "ABCanalysis" (http://cran.r-project.org/package=ABCanalysis (Ultsch & Lötsch, 2015)).

2.2.4 | Interpretation of cluster-relevant pain-related features using symbolic XAI-type machine-learning methods

In order to adhere to the concept of XAI, algorithm types were chosen that create simple recipes, usually in the form of rules or rule sets that make the class assignment transparent (Gigerenzer & Todd, 1999). Classification tree-based decision making is a traceable procedure compared to more complex classification models like random forests or deep artificial neural networks. However, growing the optimal tree is intractable even for small data sets, tree-based classification will most likely not find the optimal but rather a reasonably good solution. Therefore, it is not possible to say in advance which training algorithm will lead to the best result. In context of this study, several decision trees were used to solve a classification problem. All of these classification trees have in common that they consist of leaves representing class labels and branches representing conjunctions of observations that lead to distinct class labels. They differ in their structure and the way they are grown. For a detailed comparison of tree-based decision rules, please refer to Loh (2011, 2014). The main differences of the algorithms are summarized in Table 1.

Simple-tree–based hierarchical decision rules were determined using hierarchical classification and regression trees (CART (Breimann et al., 1993)), the recursive partitioning decision tree method, conditional inference trees (CTREE (Hothorn et al., 2006)) and decision trees created with the algorithms C4.5 (Salzberg, 1994) or C5.0 (Quinlan, 1986). In addition, non-hierarchical decision rules were created from analyses of the decision process in the random forests, using the Local Interpretable Model-Agnostic Explanations (LIME) method (Ribeiro et al., 2016) implemented in the R library "lime" (https://cran.r-project.org/package=lime (Pedersen & Benesty, 2019)) in combination with the "caret" implementation of random forests. Lime generates instance-wise explanations of classifier predictions by locally approximating the underlying model by a simple linear one. It creates a data set comprising permutated versions of the instance. The linear model is learned on this data set and weighted in favour of mistakes in perturbed instances that are close to the original one. The result is a list of individual decision rules per instance that locally interpret the classifier's prediction (Ribeiro et al., 2016).

Furthermore, non-hierarchical decision rules were generated based on partial decision trees (PART (Frank & Witten, 1998)), on repeated incremental clipping for error reduction (RIPPER (Cohen, 1995)) and on the rule-based variant of the C5.0 algorithm (Rizopoulos, 2018). As a basis for the interpretation of the recognized phenotypes of thermal pain, the rule set was chosen that most accurately assigns the subjects to the correct clusters. These calculations were performed with the R libraries "rpart" (https://cran.r-project.org/package=rpart (Therneau & Atkinson, 2019)), "party" (https://cran.r-project.org/package=party (Hothorn et al., 2006)), "RWeka" (https://cran.r-project.org/package=RWeka (Hornik et al., 2009)) and "C50" (https://CRAN.R-project.org/package=C50 (Kuhn & Quinlan, 2018)). Hyperparameter tuning included, but was not limited to, reducing the minimum number of observations that must be present in a node from the default of 20 to only 5 in Ctree or Rpart to accommodate small subgroups in the present data. In Rpart, the Gini impurity was used as splitting criterion.

In 100 runs on disjoint training and test data subsets as described above, standard measures of classification performance were used to compare the algorithms. This comprised the balanced classification accuracy and, additionally, sensitivity, specificity, precision, recall, positive and negative

predictive value (Altman & Bland, 1994a,b) and the area under the AUC-ROC. These calculations were performed with the R libraries "caret" and "pROC" (https://cran.r-project.org/package=pROC (Robin et al., 2011)). In order to control possible overfitting, all machine-learning algorithms were additionally trained with randomly permuted features. The classifier trained with these data should not perform better than guessing, i.e. should give a balanced accuracy and an AUC-ROC equal or close to 0.5. The non-parametric 95% confidence intervals (CI) of the performance measures were obtained as the 2.5th and 97.5th percentiles of the values obtained in the 100 runs.

2.2.5 | Selection of the most accurate XAI for the particular data set

The algorithms were finally ranked according to their suitability for a comprehensible cluster interpretation. Specifically, algorithms were used in 100 cross-validation runs with the full feature set and with the reduced feature set, and were trained either with the original or with randomly permuted features. Algorithms that provided better classification accuracy than guessing when trained with the permuted features were discarded due to possible overfitting. Algorithms where the 95% CI of the balanced accuracy touched or included the value of 50% were not considered because it was doubtful whether they were able to provide a valid cluster assignment. To choose the XAI for cluster interpretation, the algorithms were scored in terms of archived median balanced classification accuracy, corrected for reduction in classification accuracy from the full feature set, i.e., $Classifierscore = BA_{Reducedfeatureset}] \cdot \left(1 - (BA_{Fullfeatureset} - BA_{Reducedfeatureset}\right)$, where *BA* denotes balanced accuracy. This favours the most accurate XAI as the basis for cluster interpretation, but discriminates against the algorithm if its performance declines when trained with preselected characteristics, which would indicate that the preselection was not appropriate for the particular algorithm.

2.2.6 | Creation of a rule-based clustering with known structure

Machine-learning algorithms provided sets of rules from which the variable significance for cluster formation and, thus, cluster interpretation could be derived. However, in order to assess whether the provided rules reflect true clustering, a clustering solution was required in which these rules were known. This cannot be obtained from PCA results or k-means clustering. Therefore, a rule-based clustering was created for methodological testing purposes. The rules were derived from the present data and not arbitrarily selected, according to the following approach. First, a group structure was searched for separately in each of the d = 6 features by analyzing their modal distribution. For this purpose, their probability density distributions were described by the Pareto density estimation (PDE), which is a kernel density estimator particularly suited for group discovery (Ultsch, 2003). Modal structures were analysed by fitting Gaussian mixture models (GMM) to the PDE as $p(x) = \sum_{i=0}^{M} w_i N\left(x|m_i, s_i\right)$, where $N(x|m_i,s_i)$ denotes Gaussian probability densities with expectation values m_i and standard deviations s_i. The w_i denotes the mixture weights indicating the relative contribution of each of the M Gaussian components to the overall distribution. Models with $M = 1,...,5$ Gaussian modes were tested and the final model was selected on the basis of the Akaike Information Criterion (AIC (Akaike, 1974)) and on likelihood ratio tests (Swets, 1973) comparing the goodness of fit between the GMM with the lowest value of AIC versus the corresponding simpler model, i.e. GMM with modes *M* versus GMM with modes *M – 1*, and on visual inspection of the quantile–quantile plots of the predicted versus observed data.

The assignment of subjects to the identified subgroups was determined using Bayesian Theorem (Bayes & Price, 1763), which provides the decision limits for assigning a single observation to mode M_i based on the calculation of posterior probabilities. An automated genetic algorithm was used for this purpose as implemented in our R library "DistributionOptimization" (https://cran.r-project.org/package=DistributionOptimization (Lerch et al., 2020)). Finally, rule-based pain-related phenotypes were established by combining the one-dimensional GMM-based phenotype groups across the pain-related parameters. The Gaussian modes in the one-dimensional phenotypes were numbered in the order of increasing pain sensitivity and, thus, followed an ordinal scale with "low", "medium" or "high" pain sensitivity (Diatchenko et al., 2005). This initially resulted in 17 groups with different combinations of mode affiliations, but only groups with $n \geq 10$ members were retained, while rarer combinations were combined into a common group.

2.2.7 | Evaluations of the functioning of the cluster interpretation approach in non-pain data sets

To assess the correct functioning of the approach described above for selecting the variables relevant for clustering and choosing the XAI best suited for this purpose, selected evaluations were applied to the standard non-pain data sets. In particular, the iris data set extended by the permuted features

was subjected to feature selection to verify the reliability of the process by monitoring that the permuted features were not selected as relevant. The FCPS data set was used to readdress the need to select XAI algorithms based on their actual performance on the particular data set rather than using a predefined algorithm.

3 | RESULTS

3.1 | Participants and descriptive data

Phenotype data (pain thresholds acquired at the control side or following UV-B irradiation) were acquired from 82 subjects (age: range 19–33 years, mean ± standard deviation, *SD*: 24.7 ± 2.7 years, gender: 45 women). In the remaining subjects, the test had not taken place for all experimental conditions.

3.2 | Main results

3.2.1 | Cluster interpretation in pain thresholds to thermal stimuli

Cluster interpretation by examination of PCA and k-means clustering results

PCA revealed two PCs with eigenvalues >1 (Figure 1a), which explained 49.5% and 26.6% of the variance of the entire data set respectively. PC1 mainly carried loadings from $zHPT_{baseline}$ (25.3%), $zCPT_{baseline}$ (26.1%) and $zCPT_{UVB}$ (24.9%) and to a lesser extent to $zHPT_{UVB}$ (20.7%), while the UVB effects played almost no role (Figure 1b). PC2 mainly carried loadings from the UVB effects ($UVBEff_{heat}$: 40.4%, $UVBEff_{cold}$: 35.6%) and to a lesser degree from $zHPT_{UVB}$ (10.8%).

A cluster count of k = 2 was proposed by the largest number of indices available for this purpose. The resulting k-means clusters included $n = 47$ and $n = 35$ subjects

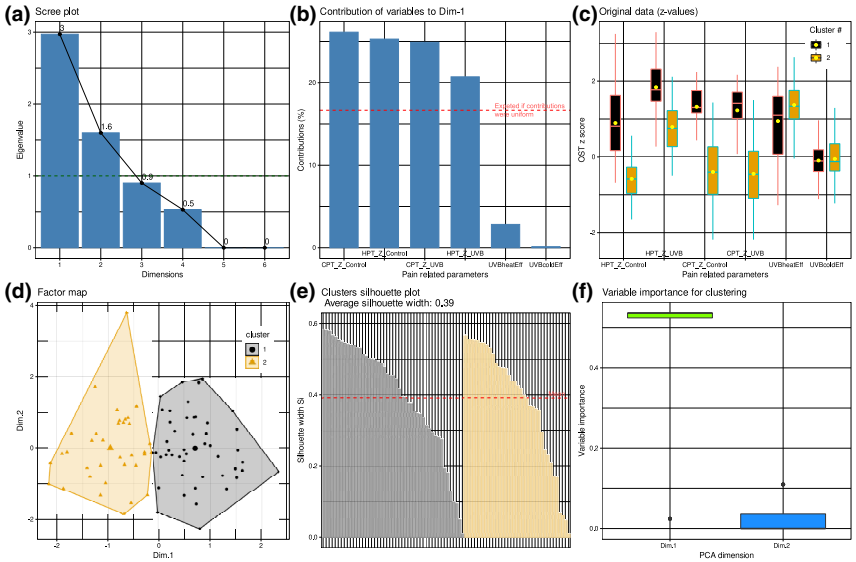


FIGURE 1 Results of principal component analysis (PCA) and k-means based clustering of the d = 6 pain-related parameters. (a) Screeplot of the amount of variation of the data captured by each PC. (b) Barplot of the contribution of each pain-related parameter to PC1 as the most relevant PC and the by far main component important to the clustering. The dashed horizontal reference line corresponds to the expected value if the contribution where uniform. (c) Box plots of original data of the pain-related parameters, shown separately for the two clusters that have resulted from the k-means–based clustering. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additional shown as yellow dots. (d) Factorial plot of the individual data points on the principal component map, obtained following k-means clustering. The coloured areas visualize the cluster separation. (e) Silhouette plot (Rousseeuw, 1987) for the two-cluster solution. Positive values indicate that the sample is away from the neighbouring cluster while negative values indicate that those samples might have been assigned to the wrong cluster because they are closer to neighbouring than to their own cluster. (f) Boxplot of the importance of data submitted to k-means clustering for the cluster solution. The values have been obtained in 100 cross-validation runs using the mean misclassification rate over all iterations as a measure of the importance of the variables. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the R packages "ggplot2" (https://cran.r-project.org/package=ggplot2) and "FactoMineR" (https://cran.r-project.org/package=FactoMineR (Le et al., 2008)). The colours were selected from the "colorblind_pal" palette provided with the R library "ggthemes" (https://cran.r-project.org/package=ggthemes (Arnold, 2019)). Variable names: HPT_Z_Control = $zHPT_{baseline}$, HPT_Z_UVB = $zHPT_{UVB}$, CPT_Z_Control = $zCPT_{baseline}$, CPT_Z_UVB = $zCPT_{UVB}$, UVBheatEff = $UVBEff_{Heat}$, UVBcoldEff = $UVBEff_{cold}$

(Figure 1c and d). The cluster solution was satisfactory, as indicated by a mean Silhouette index of 0.39 (Figure 1e). Statistical comparisons of the QST parameters between clusters by means of *t*-tests resulted in significances of $p = 2.14 \cdot 10^{-14}$ for $zHPT_{baseline}$, $p = 1.27 \cdot 10^{-9}$ for $zHPT_{UVB}$, $p = 4.07 \cdot 10^{-14}$ for $zCPT_{baseline}$, $p = 2.25 \cdot 10^{-14}$ for $zCPT_{UVB}$, $p = 0.014$ for $UVBEff_{heat}$ and $p = 0.7931$ for $UVBEff_{cold}$. After α correction, the first four remained statistically significant.

The analysis of the significance of PCs for cluster formation showed that it was mainly based on PC1 (Figure 1f). Thus, the cluster interpretation based on the PCA results led to the above-mentioned d = 4 pain-related features, which characterize the two subgroups of subjects in terms of their sensitivity to thermal pain. According to the values of the features (Figure 1c), the two clusters can be interpreted as containing either subjects with high or low thermal pain sensitivity.

Non-XAI type machine-learning–based cluster interpretation

Using d = 6 pain-related features, random forests provided a median balanced accuracy of the assignment to the two k-means clusters of 92.7% (95% CI of 100 cross-validation runs: 82.–100%; further details of performance measures obtained with full feature sets not shown). In 100 cross-validation runs on disjoint training and test data sets randomly drawn from the original data, the most frequent size of ABC subsets "A" and "B" was d = 3 items (Figure 2b). The three most frequent items in these subsets were, in decreasing order of occurrence count, $zHPT_{baseline}$, $zCPT_{UVB}$ and $zCPT_{baseline}$, whereas $zHPT_{UVB}$ *was present as frequent as* $zCPT_{baseline}$ (Figure 2a). With the same probability of importance, the choice of one of the two was appropriate. The decision could be made when 200 cross-validation runs were performed instead of 100. Then, $zHPT_{UVB}$ only achieved 62 inclusions in the relevant ABC subsets, while $zCPT_{baseline}$ was included 81 times. Training of random forests on the reduced variable space (d = 3) comprising features from subsets "A" and "B" resulted in a median balanced accuracy of 92.7% (95% CI of 100 cross-validation runs: 83.3%–100%; Table 2 and Figure 2c).

The extraction of simple combined rules from random forests using the LIME approach (RFlime) provided access to the decision on class assignment at subject level. For example, a subject belonging to cluster 1 was correctly assigned by random forests based on the conditions $-0.523 < zHPT_{baseline} \leq 0.103$ *AND* $0.952 < zCPT_{baseline} \leq 1.539$ *AND* $0.578 < zCPT_{UVB} <= 1.501$. Another subject was correctly assigned to the same class on the conditions $0.103 < zHPT_{baseline} \leq 0.937$ *AND* $0.952 < zCPT_{baseline} \leq 1.539$ *AND* $-0.235 < zCPT_{UVB} \leq 0.578$. However, the median accuracy of these class assignments was only 50% (95% CI of 100 cross-validation runs: 50%–71.4%; Table 2). Complete results of the RFlime extraction of single rules are provided in the Tables S1 and S2.

XAI-type machine-learning–based cluster interpretation

When training algorithms that create simpler hierarchical or non-hierarchical rule sets for class assignment with the d = 3 features selected by random forests and ABC analysis, C4.5 scored the best (Figure 2c and d). Its median balanced accuracy of class assignment of 90.6% (95% CI of 100 cross-validation runs: 76.5%–100%; Table 2) was only slightly lower than the accuracy achieved by random forests 92.7% (95% CI of 100 cross-validation runs: 83.3%–100%; Table 2). PART even provided a higher balanced accuracy 91.1% (95% CI of 100 cross-validation runs: 75.4%–96.9%; Table 2), but was penalized for the decrease from the performance achieved with the full feature set (Figure 2c). It is noteworthy that all algorithms delivered balanced accuracies of 50% or close to this value when trained with randomly permuted data, which supports the conclusion that the above results were not due to overfitting (Table 2).

Ability of the machine-learning–based cluster interpretation to capture rules underlying the clustering

Analysis of the probability density distribution of the d = 6 pain measures found multimodality in $zHPT_{baseline}$, $zCPT_{baseline}$, $zCPT_{UVB}$ and $UVBEff_{cold}$ (Figure 3). This was statistically supported by the lowest values of AIC (Table S3) and by the results of the likelihood ratio tests indicating statistically significantly better fits when using M modes than when using M – 1 modes, but no further improvement when using M + 1 modes (Table 3). Combination individual mode assignments based on the Bayesian boundaries initially led to 17 different vectors of each 4 modes, which after applying a cut off of 10 members led to 5 subgroups of size *n* = 24, 15, 10, 16 and 17 (for details, see Table S4). The cluster structure was based on the rules shown in Table 4.

Feature selection based on random forests and ABC analysis indicated d = 3 pain-related measures, $zHPT_{baseline}$, $zCPT_{UVB}$ and $UVBEff_{cold}$, as relevant for the cluster structure, while the fourth cluster-defining feature, $zCPT_{baseline}$, obtained the next best score but was not part of the final feature set (Figure 4a and b). With this three-feature set, random forests obtained a median balanced accuracy of class assignment of 95.2% (95% CI in 100 cross-validation runs: 66.4%–100%; Table 5), which outperformed random forests that were trained on the full variable space and resulted in a median balanced accuracy of 91.4% (95% CI in 100 cross-validation runs: 66.7%–100%). Furthermore, omitting $zCPT_{baseline}$ from the rules shown in Table 4 resulted in just *n* = 3 misclassifications, with balanced accuracies of class assignment of 93.75, 100, 97.92, 100 and 100% for subgroups #1,…,5 respectively.

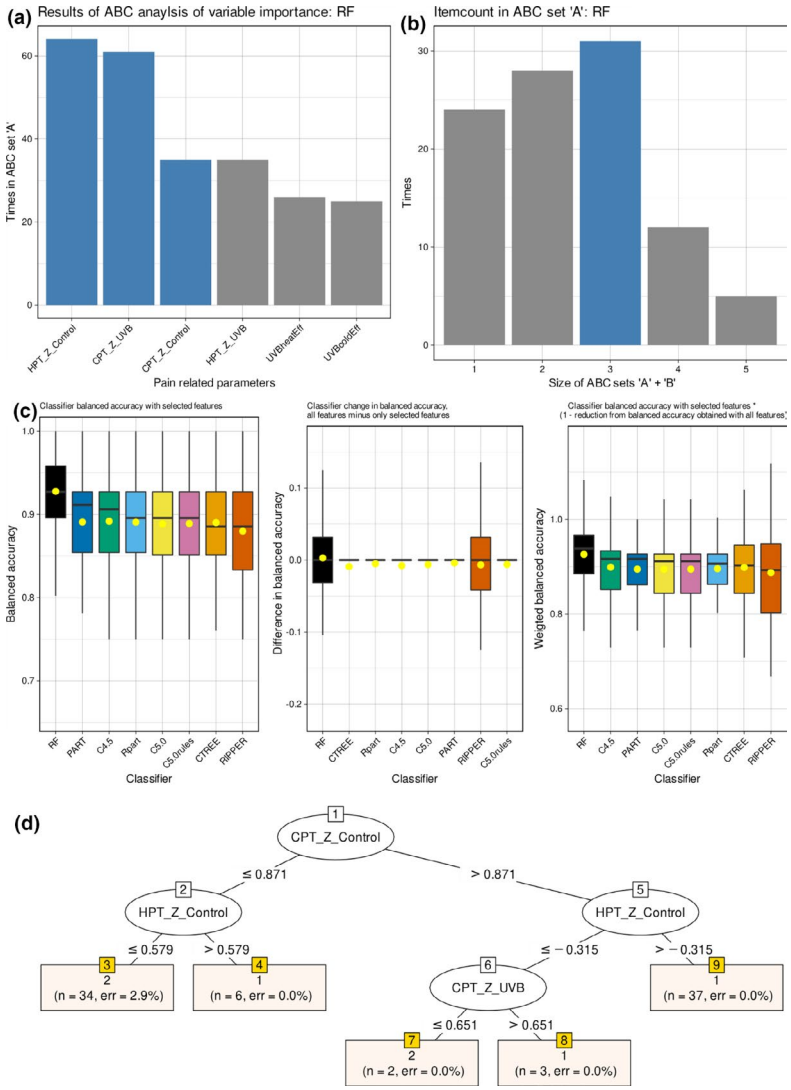


FIGURE 2 Machine-learning–based analyses of the cluster structure obtained via PCA-based projection of the data and subsequent k-means–based clustering. (a) Feature selection based on random forests, followed by computed ABC analysis. The bar graph shows the significance of the features in descending order of their occurrence in the ABC subsets "A" or "B" during 100 cross-validation runs on disjoint training and test data subsets randomly drawn from the original data set. (b) Bar plot of the size of ABC subsets. ABC subsets "A" + "B" most often had a size of d = 3 items, which caused the selection of the three most frequently included items as the final feature set important for the cluster structure. (c) Boxplots of the performance of different types of machine-learning algorithms in the assignment of subjects to the k-means–based clusters when trained with the selected features. The left panel shows the balanced accuracies of the classification obtained by the different algorithms. The middle field shows the difference in the balanced accuracy between the classification based on the full set of d = 6 pain-related features and the classification based on the selected set of d = 3 features. The right panel shows the final classifier score calculated from the values shown in the previous two panels as $Classifierscore = BA_{Reducedfeatureset} \cdot \left(1 - \left(BA_{Fullfeatureset} - BA_{Reducedfeatureset}\right)\right.$, where *BA* denotes balanced accuracy. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additionally shown as yellow dots. (d) Plot of the decision tree built by the C4.5 algorithm using the features that had passed the feature selection procedure. The panel shows the trees along with the decision limits as the basis for the assignment to either cluster. Each coloured node shows the node number (counted from top), the predicted class, the number of cases in this node and the percentage of wrongly assigned cases per node. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the R packages "ggplot2" (https://cran.r-project.org/package=ggplot2) and "RWeka" (https://cran.r-project.org/package=RWeka (Hornik et al., 2009)). The colours were selected from the "colorblind_pal" palette provided with the R library "ggthemes" (https://cran.r-project.org/package=ggthemes (Arnold, 2019)). Variable names: HPT_Z_Control = $zHPT_{baseline}$, HPT_Z_UVB = $zHPT_{UVB}$, CPT_Z_Control = $zCPT_{baseline}$, CPT_Z_UVB = $zCPT_{UVB}$, UVBheatEff = $UVBEff_{Heat}$, UVBcoldEff = $UVBEff_{cold}$. RF: random forests, RFlime: Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE: conditional inference trees, Rpart: classification and regression trees (= CART), PART: partial decision trees, RIPPER: repeated incremental clipping for error reduction

TABLE 2 Median performance measures [%] (95% CI) of 100 cross-validation runs for the correct assignment of subjects to the PCA and k-means–based clusters

	RF	RFlime	CTREE	Rpart	C4.5	C5.0	PART	RIPPER	C5.0 rules
Sensitivity, recall	96.9 (81.3–100)	100 (0–100)	100 (81.3–100)	93.8 (75–100)	93.8 (75–100)	93.8 (75–100)	93.8 (75–100)	93.8 (68.8–100)	93.8 (75–100)
Specificity	91.7 (70.6–100)	0 (0–100)	83.3 (58.3–100)	91.7 (58.3–100)	91.7 (58.3–100)	83.3 (58.3–100)	87.5 (58.3–100)	83.3 (58.3–100)	83.3 (58.3–100)
Positive predictive value, precision	93.8 (81.6–100)	57.1 (57.1–100)	88.9 (75–100)	93.3 (74.9–100)	92.9 (76.2–100)	88.9 (76.2–100)	90.6 (74.9–100)	88.9 (75–100)	88.9 (76.2–100)
Negative predictive value	96.2 (77.7–100)	45.4 (42.9–100)	100 (76.7–100)	90 (71.9–100)	91.7 (73.3–100)	91.7 (72.3–100)	91.7 (74.1–100)	91.7 (68.8–100)	91.7 (72.3–100)
F1	94.1 (84.7–100)	72.7 (40–82.1)	91.4 (81.6–100)	90.6 (80–97)	91.4 (80.6–97)	91.4 (80–97)	92.4 (80.6–97)	90.9 (78.6–98.6)	91.4 (80–97)
Balanced Accuracy	92.7 (83.3–100)	50 (50–71.4)	88.5 (76–100)	89.6 (73.9–96.9)	90.6 (76.5–96.9)	89.6 (76.5–96.9)	91.1 (75.4–96.9)	88.5 (76–98.5)	89.6 (76.5–96.9)
AUC-ROC	99.2 (94.5–100)	58.9 (50–85.7)	91.5 (76–100)	90.6 (72.8–98)	91.1 (75–96.9)	91.1 (76.1–97.7)	90.5 (75.2–96.9)	88.5 (75.9–100)	89.6 (76.5–96.9)
Balanced Accuracy permuted	47.9 (22.9–82.9)	47.9 (21.3–82.3)	53.1 (29.2–81.8)	51 (22.9–80.8)	50 (24.2–81.3)	50 (24.2–78.3)	50 (24.2–81.3)	50 (13–91.7)	50 (23.8–84)
AUC-ROC permuted	49.2 (18.2–90.4)	50.1 (9.2–95.4)	51.3 (24.1–88.1)	52.3 (19.6–87.2)	50 (26.3–78.3)	50 (21.5–84)	50 (19.2–78.3)	50 (13–91.7)	50 (24.4–84)

The models were trained on the reduced variable space as suggested by the ABC analysis (d = 3; $zHPT_{baseline}$, $zCPT_{UVB}$, and $zCPT_{baseline}$). The parameters correspond to the performance marker set implemented in the R libraries "caret" (https://cran.r-project.org/package=caret (Kuhn, 2018)) and "pROC" (https://cran.r-project.org/package=pROC (Robin et al., 2011)). RF, random forests, RFlime, Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE, conditional inference trees, Rpart, classification and regression trees (= CART), PART, partial decision trees, RIPPER, repeated incremental clipping for error reduction. With a median balanced accuracy of approximately 91%, PART and C4.5 were identified as the most accurate XAI based on the proposed ranking measure (Figure 2).

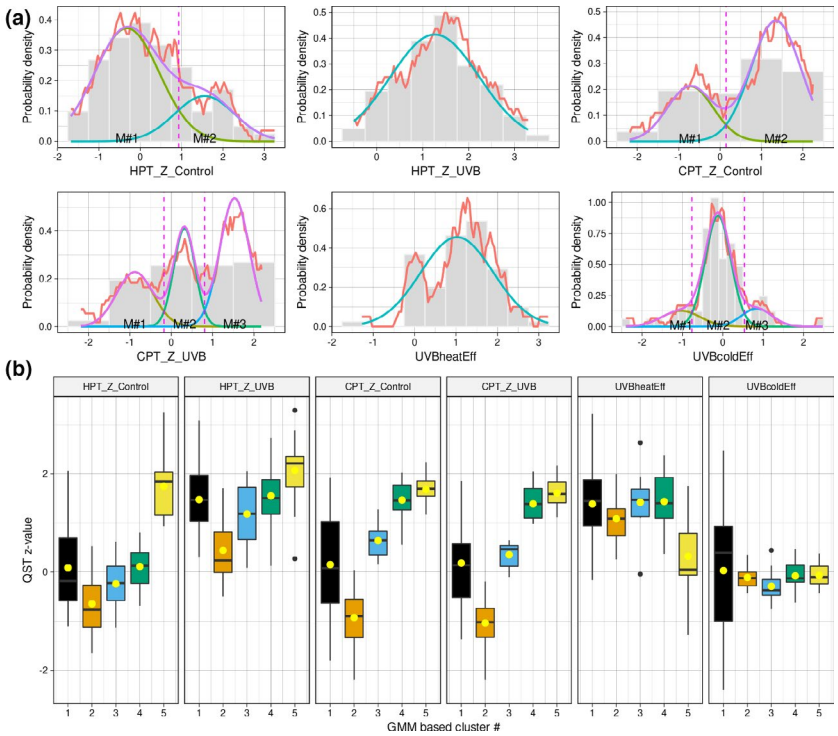


FIGURE 3 Clustering based on the modal distribution of single pain-related parameters. (a): Density plot of the QST parameters, showing the results of fitting Gaussian mixture models (GMM) to the distribution of pain thresholds to heat or cold stimuli and of the effects of UV-B irradiation on these thresholds. $M = 1,\ldots,5$ modes were tested. The final model is shown, selected based on the lowest AIC (Akaike, 1974) and on a significant likelihood ratio test indicating that it fit the data better than a model with $M - 1$ Gaussian mode. The distribution of pain-related parameters is shown as probability density function (PDF) estimated by means of the Pareto density estimation (PDE (Ultsch, 2003); black line) and overlaid on a histogram. The GMM fit is shown as a red line and the $M = 2,\ldots,5$ single mixes, if $M > 1$, are indicated as differently coloured dashed lines. The Bayesian boundaries between the Gaussians for $M > 1$ are indicated as perpendicular magenta lines. (b): Box plots showing the pattern of pain thresholds among the five different phenotypes resulting from the combination of the individual memberships to Gaussian modes 1, 2 or 3, depending on the modal distribution of thresholds to heat or cold stimuli or of the effects of UV-B irradiation on these thresholds, in the succession $zHPT_{baseline}$, $zHPT_{UVB}$, $zCPT_{baseline}$, $zCPT_{UVB}$, $UVBEff_{Heat}$ and $UVBEff_{cold}$. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additional shown as yellow dots. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the library "ggplot2" (https://cran.r-project.org/package=ggplot2 (Wickham, 2009)). An automated genetic algorithm was used implemented in our R library "DistributionOptimization" (https://cran.r-project.org/package=DistributionOptimization (Lerch et al., 2020)). Variable names: HPT_Z_Control = $zHPT_{baseline}$, HPT_Z_UVB = $zHPT_{UVB}$, CPT_Z_Control = $zCPT_{baseline}$, CPT_Z_UVB = $zCPT_{UVB}$, UVBheatEff = $UVBEff_{Heat}$, UVBcoldEff = $UVBEff_{cold}$. RF: random forests, RFlime: Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE: conditional inference trees, Rpart: classification and regression trees (= CART), PART: partial decision trees, RIPPER: repeated incremental clipping for error reduction

With a median balanced accuracy of 93.8% (95% CI in 100 cross-validation runs: 62.5%–100%) and no penalty from its performance when run with the full feature set, Rpart achieved the best score among XAI algorithms for interpreting the cluster structure (Figure 4c). The rules by which Rpart performed the cluster assignment are shown in Figure 4d and, for a more convenient comparison with the class-defining rules, in Table 4. The one-dimensional probability density distributions of the respective variables' splitting criteria as estimated from resampling experiments over 100 Rpart models (Figure 4e), as well as the splitting criteria extracted from the rules of a Rpart model that was trained on all instances (Figure 4e, dashed lines), indicate that the determined maxima mostly superimpose with the splitting criteria from the GMM-based clustering model (Figure 4e, solid lines). The only exception is a split at $UVBEff_{cold}$ at 0.24 that is indicated by a slight shoulder in the distribution.

3.2.2 | Evaluations of the cluster interpretation approach in non-pain data sets

The most relevant result from the analysis of the classical iris flower data set was that the current approach to feature selection, which was based on random forests for feature ranking and computed ABC analysis for the categorization of items into

TABLE 3 Number of Gaussian modes and parameter values of the Gaussian mixture models (GMMs) fitted to the probability density functions describing the distribution of thresholds to heat or cold stimuli or of the effects of UV-B irradiation on these thresholds

Parameter	# Modes	Gaussian #1			Gaussian #2			Gaussian #3			Bayes boundaries	
		w_1	m_1	s_1	w_2	m_2	s_2	w_3	m_3	s_3	M_1/M_2	M_2/M_3
$zHPT_{baseline}$	2	0.73	−0.319	0.78	0.27	1.554	0.722	–	–	–	0.932	–
$zHPT_{UVB}$	1	1	1.264	0.963	–	–	–	–	–	–	–	–
$zCPT_{baseline}$	2	0.296	−0.719	0.556	0.704	1.349	0.604	–	–	–	0.144	–
$zCPT_{UVB}$	3	0.253	−0.881	0.44	0.265	0.319	0.256	0.482	1.529	0.357	−0.177	0.804
$UVBEff_{Heat}$	1	1	1.024	0.877	–	–	–	–	–	–	–	–
$UVBEff_{cold}$	3	0.153	−1.035	0.475	0.71	−0.118	0.316	0.138	0.841	0.376	−0.767	0.536

Note: The GMM were implemented as $p(x) = \sum_{i=1}^{M} w_i \cdot \frac{1}{\sqrt{2\pi s_i}} \cdot e^{-\frac{(x-m_i)^2}{2s_i^2}}$, with means m_i and standard deviations s_i. The w_i denotes the mixture weights indicating the relative contribution of each Gaussian component to the overall distribution, which add up to a value of 1. Quantitative component populations can be accessed via $n_i = w_i * 82$. M denotes the number of components in the mixture.

TABLE 4 Sets of rules to assign an individual subject to the rule-based clustering

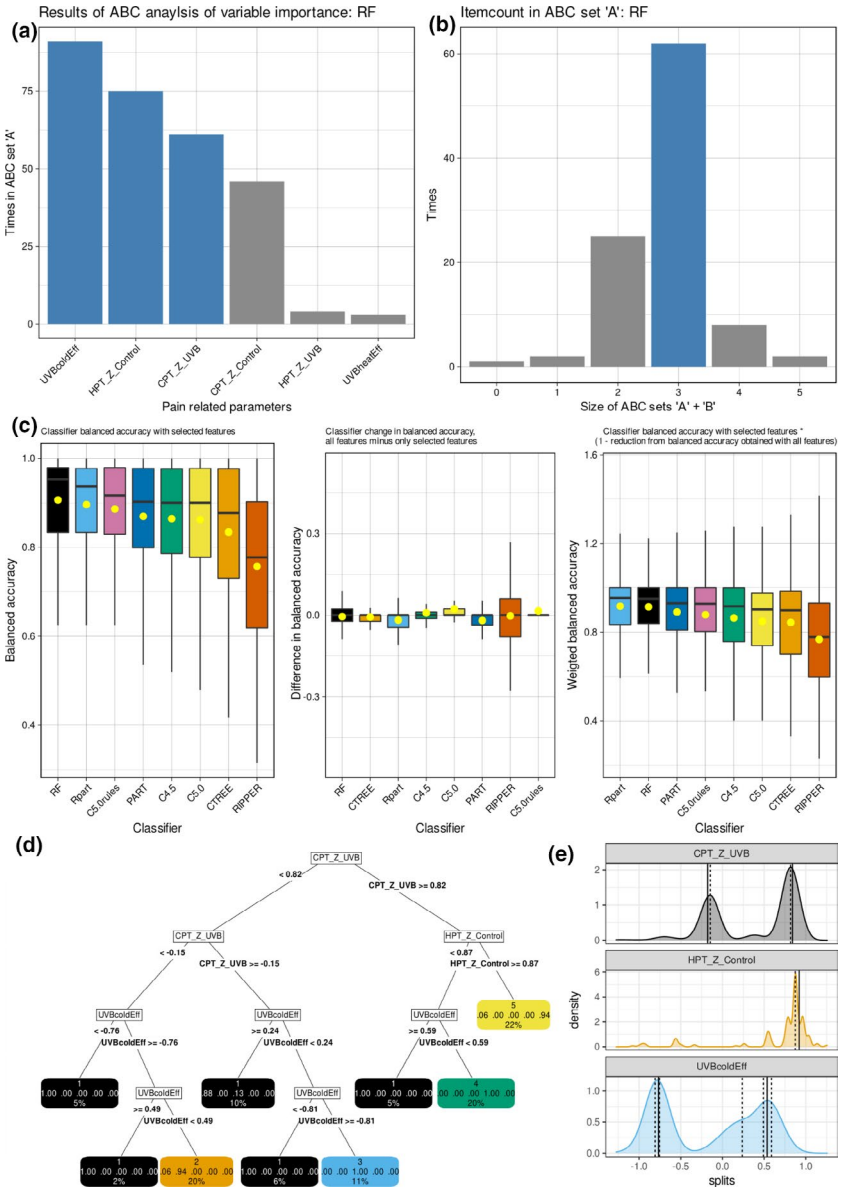
Phenotype	Rules based on GMM mode membership with known limits: Class-definitory rules	Rules based on the RPART algorithm trained with features and instance labels suggested by GMM clustering
#2	IF ($zHPT_{baseline} < 0.93$ AND $zCPT_{baseline} < 0.14$ AND $zCPT_{UVB} < -0.18$ AND $UVBEff_{cold} \geq -0.77$ AND $UVBEff_{cold} < 0.54$) THEN Group #2	IF ($zCPT_{UVB} < -0.15$ AND $UVBEff_{cold} \geq -0.76$ AND $UVBEff_{cold} < 0.49$) THEN Group #2
#3	IF ($zHPT_{baseline} < 0.93$ AND $zCPT_{baseline} \geq 0.14$ AND $zCPT_{UVB} \geq -0.18$ AND $zCPT_{UVB} < 0.8$ AND $UVBEff_{cold} \geq -0.77$ AND $UVBEff_{cold} < 0.54$) THEN Group #3	IF ($zCPT_{UVB} \geq -0.15$ AND $zCPT_{UVB} < 0.82$ AND $UVBEff_{cold} \geq -0.81$ AND $UVBEff_{cold} < 0.24$) THEN Group #3
#4	IF ($zHPT_{baseline} < 0.93$ AND $zCPT_{baseline} \geq 0.14$ AND $zCPT_{UVB} \geq 0.8$ AND $UVBEff_{cold} \geq -0.77$ AND $UVBEff_{cold} < 0.54$) THEN Group #4	IF ($zHPT_{baseline} < 0.87$ AND $zCPT_{UVB} \geq 0.82$ AND $UVBEff_{cold} < 0.59$) THEN Group #4
#5	IF ($zHPT_{baseline} \geq 0.93$ AND $zCPT_{baseline} \geq 0.14$ AND $zCPT_{UVB} \geq 0.8$ AND $UVBEff_{cold} \geq -0.77$ AND $UVBEff_{cold} < 0.54$) THEN Group #5	IF ($zHPT_{baseline} \geq 0.87$ AND $zCPT_{UVB} \geq 0.82$) THEN Group #5
#1	ELSE Group #1 i.e., all others than those captured by the four rules above	ELSE Group #1

Note: The definitory rules have derived from the Bayesian boundaries between Gaussian modes obtained from the fitted Gaussian mixture model (GMM) (Table 3). For comparison, the rules obtained with the Rpart algorithm as the best scoring XAI in the analysis of the GMM-based subgrouping are shown.

informative features to be retained or uninformative features to be dropped, did not select the permuted measures added to the data set (Figure 5a). This supports the correct functioning of the proposed approach. The algorithm selected the measure "Petal.Width" as the only relevant feature to correctly identify the iris flower species (Figure 5a and b). In fact, random forests achieved an identical balanced classification accuracy with this single feature (96.9% CI of 100 cross-validation runs: 88.2%–100%) as with the full feature set (Figure 5c). Of note, combining rules extracted by the RFlime method from the forest obtained a balanced accuracy of 87.5% in the iris data set (95% CI of 100 cross-validation runs: 68.8%–98.5%). The other algorithms achieved similarly high classification accuracies (Table 6). By contrast, the algorithms performed more heterogeneously in some of the data sets from the FCPS collection (Figure 6). Especially for the "target" data set, which comprises four classes with only three members, some of the hierarchical tree-based algorithms performed poorly. An important finding from the FCPS collection was also that random forests always and consistently performed best, near the optimum of 100% balanced accuracy.

4 | DISCUSSION

In this report, we propose an approach for the comprehensible interpretation of subgroup structures that arise in pain-related data from unsupervised analyses, such as from clustering methods that are commonly used in pain research. The approach was designed to enable clinical pain



researchers to meet the increasing demands of authorities such as the EU to communicate computer-aided decisions to the subjects they affect in an understandable way, i.e. to transfer the details of group membership from the scientific field to clinical practice. The approach included a feature selection based on variable importance for clustering and the training of supervised algorithms to generate assignment rules. We validated the approach on two different ways of clustering the same pain-related data set and on a classical PCA-based approach to data structure interpretation. We have further validated the approach of feature selection on independent data in the form of the iris flower data set, and we have validated the selection of a strong algorithm as a reference on a collection of artificial data sets. It is important to note that the present approach aims at the interpretation of data-based clusters once a cluster structure has been created. Therefore, it was not the aim of the present analysis to evaluate different clustering approaches or subgroup characteristics of pain neither to analyze the pain phenotypes in-depth; in fact, clustering only was performed because clusters were needed as a basis for the development of the interpretation approach presented in this report.

Thus, an approach for the interpretation of pain-related clusters is presented, which identifies the clustering relevant variables among the variables that have been subjected to the clustering procedure. This identification is a prerequisite for understanding the cluster structure and enables a transparent decision-making process that can tell the person concerned why he or she is placed in a particular cluster. The approach uses a combination of different types of machine learning and data science methods. A well-functioning type of machine learning algorithms is used to estimate how well a given cluster structure can be automatically captured in terms of accuracy of class assignment. The variables are

FIGURE 4 Machine-learning–based analyses of the cluster structure obtained via rules based on the Bayesian decision limits obtained by Gaussian mixture modelling of the probability density distribution of pain-related parameters. (a) Bar plot of the results of feature selection based on random forests, followed by computed ABC analysis. The graph shows the significance of the features in descending order of their occurrence in the ABC subsets "A" or "B" during 100 cross-validation runs on disjoint training and test data subsets randomly drawn from the original data set. (b) Bar plot of the size of ABC subsets "A" + "B" that most often was d = 3 items, which caused the selection of the three most frequently included items as the final feature set important for the cluster structure. (c) Boxplots of the performance of different types of machine-learning algorithms in the assignment of subjects to the rule-based clusters when trained with the selected features. The left panel shows the balanced accuracies of the classification obtained by the different algorithms. The middle field shows the difference in the balanced accuracy between the classification based on the full set of d = 6 pain-related features and the classification based on the selected set of d = 3 features. The right panel shows the final classifier score calculated from the values shown in the previous two panels as $Classifierscore = BA_{Reducedfeatureset} \cdot \left(1 - \left(BA_{Fullfeatureset} - BA_{Reducedfeatureset}\right)\right.$, where *BA* denotes balanced accuracy. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additional shown as yellow dots. (d) Decision tree built by the Rpart algorithm using the features that had passed the feature selection procedure. The panel shows the trees along with the decision limits as the basis for the assignment to either cluster. Each coloured node shows the predicted class, the predicted probability of each class, and the percentage of observations in the node. (e) Probability density distributions of the splitting criteria as determined by 100 Rpart models trained on resampled data comprising 2/3 of the original instances. The dashed vertical lines indicate the splitting criteria suggested by a Rpart model trained on the complete data set. The solid vertical lines indicate the true splitting criteria as set by the GMM-based clustering method. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the R packages "ggplot2" (https://cran.r-proje ct.org/package=ggplot2) and "rpart.plot" (https://cran.r-project.org/package=rpart.plot (Milborrow, 2018)). The colours were selected from the "colorblind_pal" palette provided with the R library "ggthemes" (https://cran.r-project.org/package=ggthemes (Arnold, 2019)). Variable names: HPT_Z_Control = $zHPT_{baseline}$, HPT_Z_UVB = $zHPT_{UVB}$, CPT_Z_Control = $zCPT_{baseline}$, CPT_Z_UVB = $zCPT_{UVB}$, UVBheatEff = $UVBEff_{Heat}$, UVBcoldEff = $UVBEff_{cold}$. RF: random forests, RFlime: Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE: conditional inference trees, Rpart: classification and regression trees (= CART), PART: partial decision trees, RIPPER: repeated incremental clipping for error reduction

then categorized in terms of their relevance for successful class assignment, and key features are automatically selected using a mathematically based item categorization technique. Multiple XAI methods are then trained on the selected features, among which one with a classification performance close to the initial algorithm is selected to extract the exact decision rules upon which the class assignment is based, provided that the XAI is capable of performing the assignment with the selected features as accurately as with all features.

The use of random forests followed by a computed ABC analysis for feature selection (Lötsch & Ultsch, 2020) proved successful in situations where the variables on which clustering depended were known, such as rule-based clustering of pain-related data, which was created specifically for this test purpose. In addition, the present approach led to a rule-generation algorithm that identified the true criteria of cluster assignment, i.e. the divisions of the variables based on Bayesian decision boundaries, with considerable accuracy. The correct functioning of the feature selection procedure could also be verified in the iris flower data set, for which others also identified the variable "Petal.Width" as the only relevant measure from which the assignment to the three species can be made correctly, except for a few single cases (Badih et al., 2019; Palczewska et al., 2014). Furthermore, a consistent observation during the analyses on pain-related and other data sets was that random forests always performed best among the machine-learning algorithms, which verified the assumed basis for their selection as the initial algorithm for feature selection.

Given the strong indications above that the proposed approach to cluster interpretation works correctly, its results obtained in a common clustering scenario of pain-related data could be considered. In the k-means clustering of the individual coordinates on the two-dimensional principal component subspace onto which the present six-dimensional pain-related data set was projected by means of a PCA, variables mainly captured in PC1 were identified as most relevant for the subsequent clustering. The present feature selection procedure succeeded in identifying these variables. Therefore, PCA and random forests followed by ABC analysis supported each other in finding the variables important for the clustering. However, the classical method failed to understand the decisions that led to the assignment of a person to a particular cluster, whereas the XAI-based approach provided rules that could be used to explain to the person concerned the exact decisions that led to the assignment to a particular cluster.

Current scientific efforts try to directly interpret sub-symbolic machine-learned algorithms without using XAIs for this task, such as the presently implemented Local Interpretable Model-Agnostic Explanations method (Ribeiro et al., 2016). The present experiments show that this indeed allows to track the decision process of cluster assignment in a fraction of the subjects, which could be used for the great clinical benefit of patients. However, this is not possible for all subjects, perhaps often not even for half of the subjects. The method succeeded better in the iris flower data set, which indicates that the modest success in the pain-related data set was not due to poor implementation. Moreover, the extracted rules are highly individual,

TABLE 5 Median performance measures [%] (95% CI) of 100 cross-validation runs for the correct assignment of subjects to the rule-based clusters

	RF	RFlime	CTREE	Rpart	C4.5	C5.0	PART	RIPPER	C5.0 rules
Sensitivity, recall	100 (33.3–100)	61.3 (0–100)	80 (0–100)	100 (33.3–100)	83.3 (22.4–100)	83.3 (22.4–100)	83.3 (25–100)	66.7 (0–100)	87.5 (33.3–100)
Specificity	95.8 (85.7–100)	95.2 (48.6–100)	95.5 (73.7–100)	95.5 (81.8–100)	95.5 (78.9–100)	95.5 (78.9–100)	95.5 (79.2–100)	95.2 (63.2–100)	95.8 (81.8–100)
Positive predictive value, precision	85.7 (50–100)	55.6 (0–100)	80 (0–100)	87.5 (40–100)	83.3 (33.3–100)	83.3 (33.3–100)	83.3 (33.3–100)	66.7 (0–100)	87.5 (38.3–100)
Negative predictive value	100 (79.2–100)	92 (70.4–100)	95.5 (76.3–100)	100 (78.7–100)	95.8 (76.3–100)	96 (76–100)	96 (76.1–100)	91.7 (70–100)	96 (82.2–100)
F1	85.7 (50–100)	59.3 (22.2–100)	80 (28.6–100)	87.5 (40–100)	80 (33.3–100)	80 (33.3–100)	80 (33.3–100)	72.7 (25–100)	85.7 (40–100)
Balanced accuracy	95.2 (66.4–100)	75 (47–100)	87.7 (46.8–100)	93.8 (62.5–100)	90 (54.9–100)	90 (55.5–100)	90.2 (57.2–100)	77.7 (45.8–100)	91.7 (59.9–100)
AUC-ROC	98.6 (94.8–100)	88.1 (75.1–94.9)	84.7 (73.3–95.7)	92.2 (82.6–99.3)	89.2 (80.7–96.6)	90.8 (83.2–98.2)	89.3 (80.3–96.2)	83 (70.3–90.8)	89 (80.3–98.8)
Balanced accuracy permuted	47.9 (30.4–77.7)	50 (28.2–75.8)	46.4 (30.4–80.9)	47.7 (30.2–76.7)	47.7 (31–76.7)	47.9 (30.4–76.3)	48.6 (30.4–75.2)	50 (37.5–64.6)	47.9 (29.5–78.6)
AUC-ROC permuted	48.5 (27.5–69.1)	50 (33.8–69)	48.7 (35.2–63.6)	49.6 (33.8–66.3)	49.8 (33.4–65.2)	48.4 (29.2–68.4)	50 (34.5–67.2)	50 (40.3–59.9)	48.6 (37.7–67.8)

Note: The models were trained on the reduced variable space as suggested by the ABC analysis (d = 3; $zHPT_{baseline}$, $zCPT_{UVB}$ and $UVBEff_{cold}$). The parameters correspond to the performance marker set implemented in the R libraries "caret" (https://cran.r-project.org/package=caret (Kuhn, 2018)) and "pROC" (https://cran.r-project.org/package=pROC (Robin et al., 2011)). RF, random forests, RFlime, Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE, conditional inference trees, Rpart, classification and regression trees (= CART), PART, partial decision trees, RIPPER, repeated incremental clipping for error reduction. With a median balanced accuracy of approximately 94%, Rpart was identified as the most accurate XAI based on the proposed ranking measure (Figure 4).

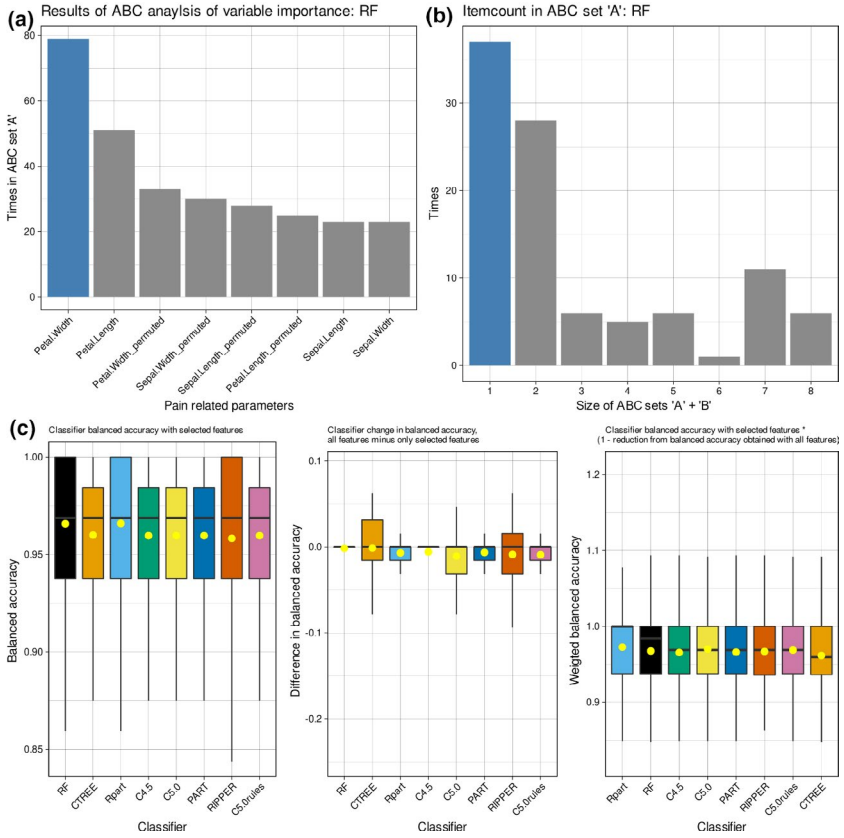


FIGURE 5 Machine-learning–based analyses of the iris flower species (Fisher, 1936). (a) Bar plot of the results of feature selection based on random forests, followed by computed ABC analysis. The graph shows the significance of the features in descending order of their occurrence in the ABC subsets "A" or "B" during 100 cross-validation runs on disjoint training and test data subsets randomly drawn from the original data set. (b) Bar plot of the size of ABC subsets "A" + "B", which most often was d = 3 items, which caused the selection of the three most frequently included items as the final feature set important for the species assignment structure. (c) Boxplots of the performance of different types of machine learning algorithms in the assignment of subjects to the rule-based clusters when trained with the selected feature. The left panel shows the balanced accuracies of the classification obtained by the different algorithms. The middle field shows the difference in the balanced accuracy between the classification based on the full set of 8 features (4 original and 4 permuted) and the classification based on the selected d = 1 feature. The right panel shows the final classifier score calculated from the values shown in the previous two panels as $Classifierscore = BA_{Reducedfeatureset} \cdot \left(1 - \left(BA_{Fullfeatureset} - BA_{Reducedfeatureset}\right)\right.$, where *BA* denotes balanced accuracy. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additional shown as yellow dots. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the R packages "ggplot2" (https://cran.r-project.org/package=ggplot2) and "rpart.plot" (https://cran.r-project.org/package=rpart.plot (Milborrow, 2018)). The colours were selected from the "colorblind_pal" palette provided with the R library "ggthemes" (https://cran.r-project.org/package=ggthemes (Arnold, 2019)). RF: random forests, RFlime: Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE: conditional inference trees, Rpart: classification and regression trees (= CART), PART: partial decision trees, RIPPER: repeated incremental clipping for error reduction

vary from subject to subject and, therefore, hardly give a clear interpretation of the characteristics of the cluster as a whole.

The present k-means–based clustering divided the cohort into two subgroups. Relevant for the subgrouping were the pain thresholds for heat and cold stimuli and additionally the individual pain thresholds for cold stimuli applied to sensitized skin. An effect of topical capsaicin sensitization on the perception of cold stimuli reproduces an observation recently made in the same laboratory (Weyer-Menkhoff & Lotsch, 2019) and in independent research units (Callsen et al., 2008; Mohr et al., 2008). The two-cluster solution separated the subjects according to either high (cluster #1) or low (cluster #2) pain sensitivity. Following the cluster interpretation from the contribution of the variables submitted to clustering, which was the aim of the present analyses, the clusters are open for further scientific investigations, e.g. on the role of genetic factors, which was analysed on another clustering in the same data set (Lötsch et al., 2020), or on gender differences. The latter were not statistically significant in the present cluster solution, as

TABLE 6 Median performance measures [%] (95% CI) of 100 cross-validation runs for the correct assignment of flowers to the iris species in the iris data set (Anderson, 1935; Fisher, 1936)

	RF	RFlime	CTREE	Rpart	C4.5	C5.0	PART	RIPPER	C5.0 rules
Sensitivity, recall	100 (81.3–100)	87.5 (43.8–100)	93.8 (81.3–100)	100 (81.3–100)	93.8 (81.3–100)	93.8 (81.3–100)	93.8 (81.3–100)	93.8 (81.3–100)	93.8 (81.3–100)
Specificity	100 (90.6–100)	100 (59.4–100)	100 (90.6–100)	100 (90.6–100)	100 (90.6–100)	100 (90.6–100)	100 (90.6–100)	100 (90.6–100)	100 (90.6–100)
Positive predictive value, precision	100 (83.3–100)	100 (55.2–100)	100 (83.3–100)	100 (83.3–100)	100 (83.3–100)	100 (83.3–100)	100 (83.3–100)	100 (82.8–100)	100 (83.3–100)
Negative predictive value	100 (91.2–100)	94.1 (76.5–100)	97 (91.4–100)	100 (91.2–100)	97 (91.3–100)	97 (91.3–100)	97 (91.3–100)	97 (91.2–100)	97 (91.3–100)
F1	96.8 (86.7–100)	81.5 (54.5–98.5)	94.1 (86.7–100)	96.8 (86.7–100)	94.1 (86.2–100)	94.1 (86.2–100)	94.1 (86.2–100)	94.1 (85.3–100)	94.1 (86.2–100)
Balanced accuracy	96.9 (89.1–100)	87.5 (68.8–98.5)	96.9 (89.8–100)	96.9 (89.1–100)	96.9 (89.1–100)	96.9 (89.1–100)	96.9 (89.1–100)	96.9 (89.1–100)	96.9 (89.1–100)
AUC-ROC	98.2 (94.9–100)	93.8 (87.7–97.6)	96.5 (92.7–99.7)	97.9 (94.2–100)	96.9 (93.7–99.5)	96.9 (94.6–99.5)	96.9 (93.7–99.5)	96.9 (93.8–99)	95.3 (92.2–99.3)
Balanced accuracy permuted	48.4 (21.9–82.1)	50 (28.1–79.8)	48.4 (23.4–78.9)	46.9 (18.8–84.4)	50 (18.8–86.8)	50 (13.8–86.8)	50 (18.8–75)	50 (20.3–75.8)	50 (19.5–86.8)
AUC-ROC permuted	52.3 (26.2–75.1)	51.7 (26.2–74.3)	51.5 (29.9–71.5)	50 (24.8–73.3)	50 (30.6–76.8)	50 (27.9–71)	50 (32.2–66.7)	50 (32.4–74.1)	50 (26.6–75)

Note: The parameters correspond to the performance marker set implemented in the R libraries "caret" (https://cran.r-project.org/package=caret (Kuhn, 2018)) and "pROC" (https://cran.r-project.org/package=pROC (Robin et al., 2011)). RF, random forests; RFlime, Local Interpretable Model-Agnostic Explanations applied on random forests; CTREE, conditional inference trees; Rpart, classification and regression trees (= CART); PART, partial decision trees; RIPPER, repeated incremental clipping for error reduction. With a median balanced accuracy of approximately 97%, all XAI performed equally.

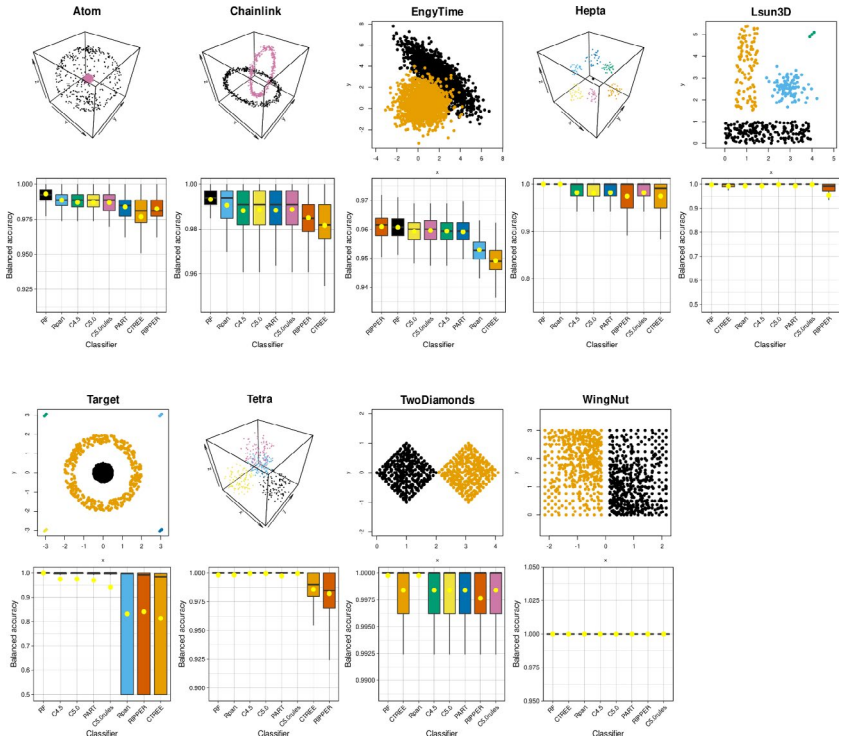


FIGURE 6 Class assignment performance of (X)AIs applied to artificial data sets from the Fundamental Clustering and Projection Suite (FCPS) Suite (Ultsch & Lötsch, 2020). Each data set is plotted, and below box plots of the classification accuracies obtained by different types of machine-learning algorithms are shown. The boxes have been constructed using the minimum, quartiles, median (solid line within the box) and maximum. The whiskers add 1.5 times the interquartile range (IQR) to the 75th percentile or subtract 1.5 times the IQR from the 25th percentile. The arithmetic mean values are additional shown as yellow dots. Please note the different scaling of the ordinates. The figure has been created using the R software package (version 4.0.2 for Linux; http://CRAN.R-project.org/ (R Development Core Team, 2008)) and the R packages "ggplot2" (https://cran.r-project.org/package=ggplot2) and "plotrix" (https://cran.r-project.org/package=plotrix (Lemon, 2006)). The colours were selected from the "colorblind_pal" palette provided with the R library "ggthemes" (https://cran.r-project.org/package=ggthemes (Arnold, 2019)). RF: random forests, RFlime: Local Interpretable Model-Agnostic Explanations applied on random forests, CTREE: conditional inference trees, Rpart: classification and regression trees (= CART), PART: partial decision trees, RIPPER: repeated incremental clipping for error reduction

a χ^2 test (Pearson, 1900) showed ($\chi^2 = 0.33639$, $df = 1$, $p = 0.5619$). Indeed, a previous analysis of an extended version of the present pain data set had identified gender differences mainly in pain induced by blunt pressure stimuli, which in an independent cohort showed the largest effect size, expressed as Cohen's d (Cohen, 1992), of the gender of the subjects among different experimental pain measures (Doehring et al., 2011).

Cluster interpretation based on the variable importance for cluster formation is an active research topic. The currently proposed approach goes beyond alternatives such as (Badih et al., 2019) by implementing XAI methods that lead directly from the identification of the relevant variables to an understandable interpretation of the cluster structure, including access to the decision process leading to the assignment of a case to a particular cluster. The method of feature selection using random forests with downstream item categorization is flexible and can be applied to many problems that can be translated into a classification problem, and has already proven successful in other biomedical context (Lötsch & Ultsch, 2020). In addition, random forests can be replaced by another method, including alternative high-performance machine-learning algorithms or other feature selection approaches such as regularization approaches implemented as, e.g. Least Absolute Shrinkage and Selection Operators (LASSO (Fonti & Belitser, 2017)).

Usually it is more difficult to interpret complex models like random forests or artificial neural networks than simpler models like decision trees or rule-based decision makers. For a better understanding, the choice of a model should, therefore, always tend towards simpler models. However, with the exception of well-structured and highly correlated data, which can sometimes be collected under laboratory conditions, it can be said that more complex models are more accurate (Arrieta et al., 2019). This is reflected in the higher performance of random forest models in pain data compared to the XAI models used. Solving this approximation dilemma is beyond the scope of this report and is the subject of current research elsewhere in algorithm development. It is approached from two sides: On the one hand, post-hoc explanatory techniques (e.g. LIME) are developed to explain complex models by simplifications. On the other hand, an attempt is made to increase the complexity

of XAI models (Arrieta et al., 2019). However, biomedical studies that compare different existing models and provide meaningful conclusions about their interpretability contribute to this research area (Tjoa & Guan, 2019). The present report contributes to this by proposing a comparative scenario for the selection of a suitable XAI, since no prior selection can be recommended, considering that the XAIs already in a limited number of data sets did not show a consistent order of classification performance.

In this context, it is perhaps worth mentioning that the presented XAI-based rules for cluster allocation have been derived from the full data set. This probably best summarizes the splits in total data set, which is the clinically relevant setting after clustering, since all subjects must be included. An alternative would be to present the model with the best performance from the cross-validation runs or the model with exactly median performance. However, this would lead to a distraction and respective tests indeed have not shown any benefit of those alternatives. In contrast, presenting all the models used in the cross-validation process would result in the XAI approach being equivalent to random forests at the end, since many trees would have to be presented. The feature selection procedure that preceded the XAI training ensured that no feature other than the selected ones, and the ones presented in the final rules set, was included during the cross-validation runs. The remaining uncertainty as to the extent to which the individual trees, i.e. trees built from a subset of data and applied to the remaining cases unseen during training, differ from the presented tree can be estimated using the experiment shown in Figure 4 E, which shows the distribution of splits during the cross-validation runs. It indicates that the central tendency of the splits reflected the truth quite well, but with considerable variance, and this uncertainty can be estimated using the classification accuracy achieved by the XAI, as shown, e.g. in Table 5. This can be reported to a patient along with the rule-based reasons for the subgroup assignments. Finally, it should be remembered that the present approach is related to clustering results, i.e. it cannot be applied to an individual patient without prior evaluating a cohort to obtain the cluster structure and assignment rules. Once this is available, a prospective patient can be assigned based on the rules, and the accuracy with which this applies to the case in question can also be communicated.

4.1 | Limitations

An intended limitation of the present report was to limit the clustering of pain data to a straight-forward approach consisting of PCA projection of the data and submission of the projected data to k-means clustering using a standard Euclidean distance. The reason for this was that for the evaluation of the present novel method, the results must be known as far as possible. In the chosen approach, the clustering was closely based on PCA and could be interpreted together with the PCA results. This would become increasingly difficult with more complex clustering methods. In fact, the data set was published previously with a clustering solution obtained with the help of emergent self-organizing maps (ESOM (Ultsch & Moerchen, 2005)) of artificial neurons, which provided a similarly good clustering (Silhouette index 0.37), but was preferred due to a comparison with hierarchical clustering (Silhouette index 0.32) and on the grounds that in several test scenarios it proved to be more reliable for clustering solutions than classical methods (Ultsch & Lötsch, 2017) because ESOM does not make such an assumption unlike, e.g. k-means, which assumes a hyperspherical form of clusters that may be inadequate. However, as mentioned above, the focus of this report was not a detailed evaluation of clustering of pain data, but the interpretation of cluster assignment once clusters have been found or alternatively an approach to create simple rules for assigning future subjects to relevant subgroups.

A second limitation of the report is the lack of comparative benchmark experiments of different machine-learning methods used for the feature selection, which was mentioned above as a sign of the flexibility of the present approach. The reason for this was that random forests worked very well in pain-related and other data sets with achieved median balanced accuracies in a range of 92.7% to 96.9%. It should be noted, however, that this may not always be the case, and since the present approach, by translating the cluster interpretation into a classification problem, relies on excellent classifier performance, the inclusion of alternative types of machine-learning algorithms may be triggered when random forests provide unsatisfactory results. Therefore, alternative feature selection methods such as the regression-based LASSO (Fonti & Belitser, 2017) or an alternative random-forests based known as "Boruta" (Kursa & Rudnicki, 2010) were not included.

Furthermore, this report did not include all types of XAI listed in the overview given in Table 1. The reason for this was the lack of R implementations of some methods. While the extraction of combined rules using the LIME approach to test it against the other algorithms (“RFlime” columns in Tables 2,5 and 6) was implemented for the purposes of this report, the implementation of additional algorithms in novel R libraries was considered to be beyond the scope of the present analyses. The use of alternative software was not considered as an option because differences in the software packages would have distracted the focus from comparing the algorithms by requiring.

Finally, the focus was on the variable importance for cluster interpretation after the establishment of clusters. The clustering structure itself was chosen as a basic k-means algorithm, as discussed above. Whether the two clusters represent a general finding or only characterize the current patient group

could not be investigated due to the lack of a similar data set, i.e. data collected in exactly the same way as in the present analysis. The number of samples was considered insufficient for a split into two sets, as this would have jeopardized the validation of the current cluster interpretation, which involved massive resampling procedures that already split the data set into disjoint subsets. Nevertheless, the finding of subjects with a tendency to high or low pain sensitivity seems to be undisputed and has often been reported as a subgroup structure in pain-related data (Diatchenko et al., 2005).

4.2 | Transfer learning remarks

The present report proposes a bioinformatics approach to allocate clusters/subgroups transparently; it is therefore methodologically located in data science, while thematically it is located in pain research. This is a typical setting in current methods of data science, which increasingly allow for data-driven research approaches and the extraction of information and the generation of knowledge from these data, and which are predominantly multidisciplinary, cross-agency, cross-sector and cooperative (President's Information Technology Advisory, 2005). In this report, this has been exploited by applying the proposed computational approach to pain-related data together with non-pain–related data. Complex high-dimensional pain data often have the disadvantage that an eventual subgroup structure is unknown, except for simple settings such as patients with pain versus pain-free subjects where the present approach would be unnecessary. However, if the subgroup structure is determined on the basis of the data and not by following predefined clinical definitions, the full truth may not be known.

It is a standard procedure in data science to test the correct functioning of an approach on data sets where the ground truth is known. Hence, the current use of the iris flower data set, where the group structure consists of three biological defined iris species. This data set was used for method development in statistics when Fisher introduced linear discriminant analysis (Fisher, 1936), and it has been widely used in computer science for testing algorithms. Here, it has been used to assess whether the proposed method can select the relevant variables that allow for subgroup assignment. The FCPS collection (Ultsch & Lötsch, 2020) was developed for testing clustering methods and, therefore, provided several data sets with a defined subgroup structure to evaluate the comparative performance of machine-learning algorithms for group assignment. The FCPS results support the selection of random forests as a strong classifier, which serves as a reference for the other classifiers with regard to the maximum achievable performance of subgroup allocation. On the one hand, these data sets provide a reference for XAI by allowing weak classification performance to be attributed to weaknesses in the XAI implementation or weaknesses in the data set. On the other hand, these data sets are not pain related, so it was necessary to develop the method in a real pain data set to ensure that the crossing of the research areas was successful.

Taken together, computational approaches cross the border of a particular research area and require collaborative knowledge of both areas, the topic itself to judge whether the obtained results are plausible and the methods assuring technical correct results. The authors propose that the final explanation of a cluster structure may be best placed with the topical pain expert who can judge its clinical relevance, as long as the selection does not collide with the data analysis results.

5 | CONCLUSIONS

An XAI-based approach to interpreting cluster structures found in pain-related data is proposed, based on assessments of variable importance. The approach uses different types of machine-learning and data science methods to (1) identify the relevant features on which the cluster structure is based and (2) track the precise decision-making process of cluster assignment. It uses high-performance machine-learning algorithms for feature selection and passes the results to an XAI algorithm to generate understandable class assignment rules. A score is proposed for the XAI selection that considers both the absolute cluster allocation performance and the algorithm's ability to achieve this maximum with the pre-selected variables. The approach was compared with the interpretation of cluster structures based on PCA results as a common procedure in pain research. Further experiments with standard data sets with known class structures emphasized the correct functioning of the selection of important variables on which the clustering is based, and the selection of random forests machine learning as a consistently well-performing type of machine-learning algorithms. The proposed approach to the interpretation of clusters in pain-related data makes it possible to follow the Council of the European Union's view that computer-assisted decisions must be transparent so that they can be communicated to affected patients in an understandable way (Hamon et al., 2020). Finally, it is stressed that it is crucial for cluster interpretation to understand clustering from the importance of the variables used in this process before the relevance of other factors not used in cluster identification can be analysed in the context of the clustering.

6 | SUPPLEMENTARY INFORMATION

Supplementary information includes a detailed description of the rule-generating XAI algorithms used in the

present analyses (XAI_SI_05.pdf) and 4 Supplementary Tables displaying (1) the case-wise rules for the assignment to k-means–based clusters extracted from random forests using the LIME approach (Table S1; file "RFlime_PCAkmeansClusters_Casewise.xlsx") and (2) the combined rules derived from that analysis (Table S2; RFlime_PCAkmeansClusters_CombinedRules.xlsx), (3) statistical details of the Gaussian mixture modelling (Table S1 Table 1; file "SupplementaryTables.docx") and (4) details of the subjects grouping according to Gaussian mode membership of the pain data (Table S2; file "SupplementaryTables.docx").

ACKNOWLEDGEMENTS

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors have declared that no competing interests exist.

ORCID

Jörn Lötsch https://orcid.org/0000-0002-5818-6958
Sebastian Malkusch https://orcid.org/0000-0001-6766-140X

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Altman, D. G., & Bland, J. M. (1994a). Diagnostic tests 2: Predictive values. *BMJ*, *309*, 102.

Altman, D. G., & Bland, J. M. (1994b). Diagnostic tests. 1: Sensitivity and specificity. *BMJ*, *308*, 1552.

Anderson, E. (1935). The irises of the Gaspé peninsula. *Bulletin of the American Iris Society*, *59*, 2–5.

Arnold, J. B. (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'.

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. et al (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.

Badih, G., Pierre, M., & Laurent, B. (2019). Assessing variable importance in clustering: A new method based on unsupervised binary decision trees. *Computational Statistics*, *34*, 301–321. https://doi.org/10.1007/s00180-018-0857-0

Bayes, M., & Price, M. (1763). An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions 53*, 370–418.

Bezdek, J. C., Keller, J. M., Krishnapuram, R., Kuncheva, L. I., & Pal, N. R. (1999). Will the real iris data please stand up? *IEEE Transactions on Fuzzy Systems*, *7*, 368–369. https://doi.org/10.1109/91.771092

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commerciali Di Firenze*, *8*, 3–62.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Breimann, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1993). *Classification and regression trees*. Chapman and Hall.

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The Balanced Accuracy and Its Posterior Distribution. In Pattern Recognition (ICPR), 2010 20th International Conference on, pp. 3121–3124.

Callsen, M. G., Moller, A. T., Sorensen, K., Jensen, T. S., & Finnerup, N. B. (2008). Cold hyposensitivity after topical application of capsaicin in humans. *Experimental Brain Research*, *191*, 447–452. https://doi.org/10.1007/s00221-008-1535-1

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen, W. W. (1995). Fast effective rule induction. In ICML.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, *56*, 64–73. https://doi.org/10.1145/2500499

Diatchenko, L., Slade, G. D., Nackley, A. G., Bhalang, K., Sigurdsson, A., Belfer, I., Goldman, D., Xu, K. E., Shabalina, S. A., Shagin, D., Max, M. B., Makarov, S. S., & Maixner, W. (2005). Genetic basis for individual variations in pain perception and the development of a chronic pain condition. *Human Molecular Genetics*, *14*, 135–143. https://doi.org/10.1093/hmg/ddi013

Dimova, V., Oertel, B. G., & Lötsch, J. (2016). Using a standardized clinical quantitative sensory testing battery to judge the clinical relevance of sensory differences between adjacent body areas. *Clinical Journal of Pain*. *33*(1), 37–43.

Doehring, A., Küsener, N., Flühr, K., Neddermeyer, T. J., Schneider, G., & Lötsch, J. (2011). Effect sizes in experimental pain produced by gender, genetic variants and sensitization procedures. *PLoS One*, *6*, e17724. https://doi.org/10.1371/journal.pone.0017724

Fillingim, R. B., Bruehl, S., Dworkin, R. H., Dworkin, S. F., Loeser, J. D., Turk, D. C., Widerstrom-Noga, E., Arnold, L., Bennett, R., Edwards, R. R., Freeman, R., Gewandter, J., Hertz, S., Hochberg, M., Krane, E., Mantyh, P. W., Markman, J., Neogi, T., Ohrbach, R., … Wesselmann, U. (2014). The ACTTION-American Pain Society Pain Taxonomy (AAPT): An evidence-based and multidimensional approach to classifying chronic pain conditions. *The Journal of Pain*, *15*, 241–249. https://doi.org/10.1016/j.jpain.2014.01.004

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam Research Paper in Business Analytics*, *30*, 1–25.

Frank, E., & Witten, I. H. (1998). Generating Accurate Rule Sets Without Global Optimization. In ICML.

Gigerenzer, G., Todd, P. M., & The ABC Research Group, Evolution and cognition (1999). Fast and frugal heuristics: The adaptive toolbox. In G Gigerenzer & P.M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press).

Gustorff, B., Anzenhofer, S., Sycha, T., Lehr, S., & Kress, H. G. (2004). The sunburn pain model: The stability of primary and secondary hyperalgesia over 10 hours in a crossover setting. *Anesthesia & Analgesia*, *98*, 173–177. table of contents. https://doi.org/10.1213/01.ane.0000093224.77281.a5.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika*, *19*, 149–161. https://doi.org/10.1007/BF02289162

Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and Explainability of Artificial Intelligence - From technical to policy solutions. (Luxembourg, Publications Office of the European Union, Luxembourg).

Harrison, G. I., Young, A. R., & McMahon, S. B. (2004). Ultraviolet radiation-induced inflammation as a model for cutaneous hyperalgesia. *The Journal of Investigative Dermatology*, *122*, 183–189. https://doi.org/10.1046/j.0022-202X.2003.22119.x

Ho, T. K. (1995). Random decision forests. In Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1 (IEEE Computer Society), p. 278.

Hoffmann, R. T., & Schmelz, M. (1999). Time course of UVA- and UVB-induced inflammation and hyperalgesia in human skin. *European Journal of Pain*, *3*, 131–139. https://doi.org/10.1053/eujp.1998.0106

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–90.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. *Comput Stat*, *24*, 225–232. https://doi.org/10.1007/s00180-008-0119-7

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 498–520. https://doi.org/10.1037/h0070888

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*, 651–674. https://doi.org/10.1198/106186006X133933

Juran, J. M. (1975). The non-pareto principle; Mea culpa. *Quality Progress*, *8*, 8–9.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187–200. https://doi.org/10.1007/BF02289233

Kaiser, U., Kopkow, C., Deckert, S., Neustadt, K., Jacobi, L., Cameron, P., De Angelis, V., Apfelbacher, C., Arnold, B., Birch, J., Bjarnegård, A., Christiansen, S., C de C Williams, A., Gossrau, G., Heinks, A., Hüppe, M., Kiers, H., Kleinert, U., Martelletti, P., … Schmitt, J. (2018). Developing a core outcome domain set to assessing effectiveness of interdisciplinary multimodal pain therapy: The VAPAIN consensus statement on core outcome domains. *Pain*, *159*, 673–683. https://doi.org/10.1097/j.pain.0000000000001129

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, *29*, 119–127. https://doi.org/10.2307/2986296

Kim, H., & Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, *96*, 589–604. https://doi.org/10.1198/016214501753168271

Kuhn, M. (2018). caret: Classification and Regression Training.

Kuhn, M., & Quinlan, R. (2018). C50: C5.0 decision trees and rule-based models.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *Journal of Statistical Software*, *36*, 13.

Le, S., Josse, J., & Husson, F. C.(2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*, 1–18.

Leisch, F. (2006). A toolbox for K-centroids cluster analysis. *Computational Statistics & Data Analysis*, *51*, 526–544.

Lemon, J. (2006). Plotrix: A package in the red light district of R. *R-News*, *6*, 8–12.

Lerch, F., Ultsch, A., & Lotsch, J. (2020). Distribution Optimization: An evolutionary algorithm to separate Gaussian mixtures. *Scientific Reports*, *10*, 648. https://doi.org/10.1038/s41598-020-57432-w

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*, 18–22.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Ann Appl Stat*, *3*, 1710–1737. https://doi.org/10.1214/09-AOAS260

Loh, W.-Y. (2011). Classification and regression trees. *Wires Data Mining and Knowledge Discovery*, *1*, 14–23. https://doi.org/10.1002/widm.8

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review*, *82*, 329–348. https://doi.org/10.1111/insr.12016

Loh, W.-Y., & Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, *7*, 815–840.

Loh, W.-Y., & Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis. *Journal of the American Statistical Association*, *83*, 715–725. https://doi.org/10.1080/01621459.1988.10478652

Lötsch, J., Geisslinger, G., Heinemann, S., Lerch, F., Oertel, B. G., & Ultsch, A. (2017). Quantitative sensory testing response patterns to capsaicin- and UV-B-induced local skin hypersensitization in healthy subjects: A machine-learned analysis. *Pain*, *159*, 11–24.

Lötsch, J., Kringel, D., Geisslinger, G., Oertel, B. G., Resch, E., & Malkusch, S. (2020). Machine-learned association of next-generation sequencing-derived variants in thermosensitive ion channels genes with human thermal pain sensitivity phenotypes. *International Journal of Molecular Sciences*, *21*, 4367. https://doi.org/10.3390/ijms21124367

Lötsch, J., & Ultsch, A. (2019). Current projection methods-induced biases at subgroup detection for machine-learning based data-analysis of biomedical data. *International Journal of Molecular Sciences*, *21*, 79 https://doi.org/10.3390/ijms21010079

Lötsch, J., & Ultsch, A. (2020). Random forests followed by computed ABC analysis as a feature selection method for machine-learning in biomedical data. In T. Imaizumi, A. Okada, S. Miyamoto, F. Sakaori, Y. Yamamoto, & M. Vichi (Eds.), *Advanced Studies in Classification and Data Science*, Singapore: Springer. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-981-15-3311-2_5.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability*, Volume *1*: Statistics (pp. 281–297), University of California Press.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2017). cluster: Cluster analysis basics and extensions.

Magerl, W., Krumova, E. K., Baron, R., Tölle, T., Treede, R. D., & Maier, C. (2010). Reference data for quantitative sensory testing (QST): Refined stratification for age and a novel method for statistical comparison of group data. *Pain*, *151*, 598–605. https://doi.org/10.1016/j.pain.2010.07.026

Milborrow, S. (2018). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'.

Mohr, C., Leyendecker, S., Mangels, I., Machner, B., Sander, T., & Helmchen, C. (2008). Central representation of cold-evoked pain relief in capsaicin induced pain: An event-related fMRI study. *Pain*, *139*, 416–430. https://doi.org/10.1016/j.pain.2008.05.020

Murphy, K. P. (2012). *Machine learning*. A Probabilistic Perspective (The MIT Press).

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19*, 113–126. https://doi.org/10.1145/360018.360022

Palczewska, A., Palczewski, J., Marchese Robinson, R., & Neagu, D. (2014). Interpreting random forest classification models using a feature contribution method. In T. Bouabana-Tebibel & S. H. Rubin (Eds.), *Integration of reusable systems* (pp. 193–218). Springer International Publishing.

Pareto, V. (1909). Manuale di economia politica, Milan: Società editrice libraria, revised and translated into French as Manuel d'économie politique.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series*, *5*(50), 157–175.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*, 559–572.

Pedersen, T. L., & Benesty, M. (2019). lime: Local Interpretable Model-Agnostic Explanations.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*, 171–212. https://doi.org/10.1109/TIT.1954.1057460

Pfaffel, O. (2020). FeatureImpCluster: Feature Importance for Partitional Clustering.

President's Information Technology Advisory, C. (2005). Report to the president: computational science: Ensuring America's Competitiveness.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106.

Quinlan, J. R. (2014). C4.5 : Programs for machine learning.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing.

Ribeiro, M. T., 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 (San Francisco, California, USA, Association for Computing Machinery), pp. 1135–1144.

Rizopoulos, D. (2018). Max Kuhn and Kjell Johnson. applied predictive modeling. New York, Springer. *Biometrics*, *74*, 383.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 77.

Rolke, R., Baron, R., Maier, C., Tolle, T. R., Treede, R. D., Beyer, A., Binder, A., Birbaumer, N., Birklein, F., Botefur, I. C. et al (2006). Quantitative sensory testing in the German Research Network on Neuropathic Pain (DFNS): Standardized protocol and reference values. *Pain*, *123*, 231–243.

Rolke, R., Magerl, W., Campbell, K. A., Schalber, C., Caspari, S., Birklein, F., & Treede, R. D. (2006). Quantitative sensory testing: A comprehensive protocol for clinical trials. *European Journal of Pain*, *10*, 77–88.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Salzberg, S. L. (1994). C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers Inc, 1993. *Machine Learning*, *16*, 235–240. https://doi.org/10.1007/BF00993309

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, *19*(2), 279–281.

Smolensky, P. (2010). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*, 1–23. https://doi.org/10.1017/S0140525X00052432

Student, (1908). Probable Error of a Correlation Coefficient Student. *Biometrika*, *6*. 302–310. https://doi.org/10.1093/biomet/6.2-3.302.

Swets, J. A. (1973). The relative operating characteristic in psychology: A technique for isolating effects of response bias finds wide use in the study of perception and cognition. *Science*, *182*, 990–1000. https://doi.org/10.1126/science.182.4116.990

Therneau, T., & Atkinson, B. (2019). rpart: Recursive Partitioning and Regression Trees.

Tjoa, E., & Guan, C. (2019). *A Survey on Explainable Artificial Intelligence (XAI)*. Towards Medical XAI.

Ultsch, A. (2003). Pareto density estimation: A density estimation for knowledge discovery. In D. Baier & K. D. Werrnecke, (Eds.), *Innovations in Classification, Data Science, and Information Systems - Proceedings 27th Annual Conference of the German Classification Society (GfKL)*, (621–628). Springer.

Ultsch, A., & Lötsch, J. (2015). Computed ABC analysis for rational selection of most informative variables in multivariate data. *PLoS One*, *10*, e0129767. https://doi.org/10.1371/journal.pone.0129767

Ultsch, A., & Lötsch, J. (2017). Machine-learned cluster identification in high-dimensional data. *Journal of Biomedical Informatics*, *66*, 95–104. https://doi.org/10.1016/j.jbi.2016.12.011

Ultsch, A., & Lötsch, J. (2020). The fundamental clustering and projection Suite (FCPS): A dataset collection to test the performance of clustering and data projection algorithms. *Data*, *5*, 13. https://doi.org/10.3390/data5010013

Ultsch, A., & Moerchen, F. (2005). ESOM-Maps: Tools for clustering, visualization, and classification with Emergent SOM. Technical Report Dept of Mathematics and Computer Science. Germany: University of Marburg.

Vartiainen, P., Heiskanen, T., Sintonen, H., Roine, R. P., & Kalso, E. (2016). Health-related quality of life and burden of disease in chronic pain measured with the 15D instrument. *Pain*, *157*, 2269–2276. https://doi.org/10.1097/j.pain.0000000000000641

Vollert, J., Mainka, T., Baron, R., Enax-Krumova, E. K., Hüllemann, P., Maier, C., Pfau, D. B., Tölle, T., & Treede, R. D. (2015). Quality assurance for Quantitative Sensory Testing laboratories: Development and validation of an automated evaluation tool for the analysis of declared healthy samples. *Pain*, *156*, 2423–2430. https://doi.org/10.1097/j.pain.0000000000000300

Weyer-Menkhoff, I., & Lotsch, J. (2019). TRPA1 sensitization produces hyperalgesia to heat but not to cold stimuli in human volunteers. *Clinical Journal of Pain*. https://doi.org/10.1097/AJP.0000000000000677

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32:AID-CNCR2820030106>3.0.CO;2-3

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Lötsch J, Malkusch S. Interpretation of cluster structures in pain-related phenotype data using explainable artificial intelligence (XAI). *Eur J Pain*. 2021;25:442–465. https://doi.org/10.1002/ejp.1683