Evaluation of Human-Understandability of Global Model Explanations Using Decision Tree

Adarsa Sivaprasad[1(✉)], Ehud Reiter[1], Nava Tintarev[2], and Nir Oren[1]

[1] Department of Computing Science, University of Aberdeen, Aberdeen, UK
{.sivaprasad.22,e.reiter,n.oren}@abdn.ac.uk
[2] Maastricht University, Maastricht, Netherlands
n.tintarev@maastrichtuniversity.nl

Abstract. In explainable artificial intelligence (XAI) research, the predominant focus has been on interpreting models for experts and practitioners. Model agnostic and local explanation approaches are deemed interpretable and sufficient in many applications. However, in domains like healthcare, where end users are patients without AI or domain expertise, there is an urgent need for model explanations that are more comprehensible and instil trust in the model's operations. We hypothesise that generating model explanations that are narrative, patient-specific and *global* (holistic of the model) would enable better understandability and enable decision-making. We test this using a decision tree model to generate both local and global explanations for patients identified as having a high risk of coronary heart disease. These explanations are presented to non-expert users. We find a strong individual preference for a specific type of explanation. The majority of participants prefer global explanations, while a smaller group prefers local explanations. A task based evaluation of mental models of these participants provide valuable feedback to enhance narrative global explanations. This, in turn, guides the design of health informatics systems that are both trustworthy and actionable.

Keywords: Global Explanation · End-user Understandability · Health Informatics

1 Introduction

The field of explainable artificial intelligence (XAI) has witnessed significant advancements, primarily focusing on the interpretability of models. However, the interpretability of an AI model for developers does not seamlessly translate into end-user interpretability [3]. Even inherently interpretable models like decision trees (DT) and decision lists are challenging to use in applications due to the complexity and scale of data. Hence popular explanation techniques interpret black box models by considering an individual input and corresponding prediction - *local explanations*. Model-agnostic explanations such as Shapley values

S. Nowaczyk et al. (Eds.): ECAI 2023 Workshops, CCIS 1947, pp. 43–65, 2024.
https://doi.org/10.1007/978-3-031-50396-2_3

and Local Interpretable Model-Agnostic Explanations (LIME) offer insights into the features contributing to an individual prediction, revealing the importance of specific characteristics in decision-making. Nevertheless, they do not capture the complete model functioning, comprehensive utilization of data, and, most importantly, the interactions among features. They lack the ability to facilitate generalization or provide a complete mental model of the system's workings.

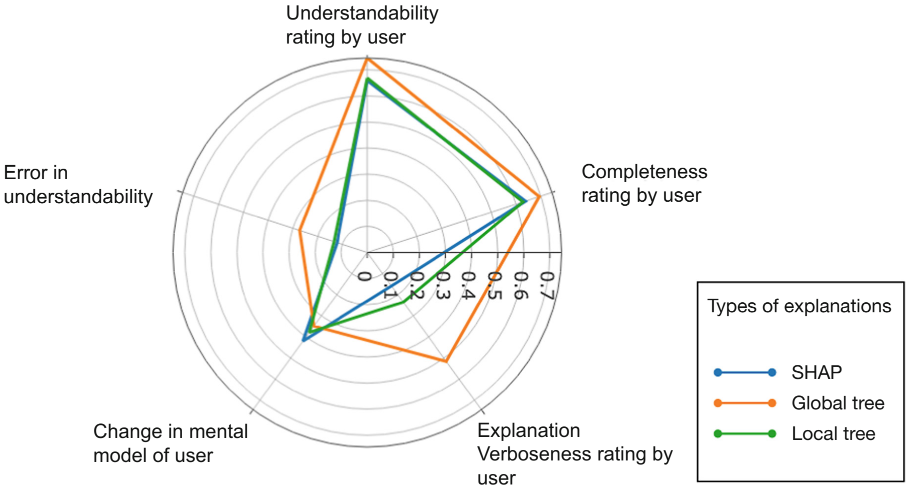


Fig. 1. A comparison of Local SHAP, Local and Global tree explanation of CHD risk prediction using decision tree model. Different evaluation parameters are computed based on end-user feedback of the explanation.

In critical domains such as healthcare and financial predictions, the interpretability of AI models by end-users holds significant importance. The understandability of the underlying AI model and the trust in its predictions can have life-altering implications for stakeholders. Enabling user intervention and action to modify predicted outcomes require explanations that address the *How* and *Why* questions, as well as convey causal relationships [18,21]. Achieving this necessitates an overall comprehension of the model. Further, the explanation should not only align the user's mental model with the AI system's model but also be perceived as understandable and trustworthy. We propose that a global model explanation hold greater potential for providing understandability and building trust compared to local model explanations. This study is a preliminary step towards testing this.

What qualifies as a global explanation and what methodologies would provide an overall understandability is relatively less researched. The comparison between global model explanations and local explanations for end users, along with various presentation aspects such as narrative and visualization, bears significance when building explanation-centric applications. This study delves into the understandability of local and global explanations, specifically in the context of a coronary heart prediction model. We address the following research question:

1. For non-expert users, do global explanations provide a better understanding of the AI's reasoning in comparison to (only) local explanations?
2. As the complexity of the explanation increases is there a difference in understandability and user preference for local and global explanations?

We use decision tree (DT) models which are interpretable by design, and construct local and global explanations with varying levels of complexity. We gauge the perceived understandability of these models and evaluate their effectiveness based on predefined tasks. We also measure the changes in users' mental models following exposure to the explanations. Figure 1 shows different evaluation parameters. The experiment identifies preferences in explanation types among different participant groups. It is found that while complexity does not have a significant effect on perceived understandability and completeness of explanation, errors in understanding increase with complexity. The obtained results offer valuable insights for designing narrative explanations for end-users and highlight the majority of participant preference for global explanations in healthcare risk models.

2 Related Work

In healthcare, a risk score is a quantifiable measure to predict aspects of a patient's care such as morbidity, the chance of response to treatment, cost of hospitalisation etc. Risk scoring is utilised for its predictive capabilities and in managing healthcare at scale. A predicted risk score is representative of the probability of an adverse outcome and the magnitude of its consequence. Article 22 of the General Data Protection Regulation (GDPR) mandates human involvement in automated decision-making and in turn understandability of a risk prediction model. Hence the use of risk scores requires the effective communication of these scores to all stakeholders - doctors, patients, hospital management, health regulators, insurance providers etc. With statistical and black-box AI models used in risk score computations, this is an added responsibility of the AI model developer to ensure the interpretability of these systems to all stakeholders.

Current regulations such as model fact tables [25] are useful for clinicians and approaches of local model interpretation [15,24] to model developers. For a non-expert end-user who has limited domain knowledge and who is not trained to understand a fact table, these approaches will not explain a recommendation given to them. Further, explaining a risk prediction model to the end user should address the perceived risk from numeric values and previous knowledge of the user, any preferences and biases. In other words, the explanation presentation should address socio-linguistic aspects [18] involved.

Researchers have recognized that a good explanation should aim to align the user's mental model with the mental model of the system, promoting faithful comprehension and reducing cognitive dissonance [18]. Achieving such effectiveness is very context-dependent [1]. However, aspects of explanation presentation generalise across a broad spectrum of applications. The significance of narrative-style explanations is emphasised by [23] while [26], highlights the effectiveness of a combined visual and narrative explanation. Recent studies have evaluated existing systems in use [6,16] and calls for focus on the design choices for explanation presentation in health informatics. Further, with tools available in the public domain such as QRisk[1] from National Health Service (NHS), evaluating the impact and actionability of explanation approaches in use would enable improving them and ensure their safe usage.

Before looking into evaluating black-box models, it would be worthwhile to explore what constitutes a good explanation in interpretable models such as DTs, decision lists [13] etc. DT algorithms are methods of approximating a discrete-valued target by recursively splitting the data into smaller subsets based on the features that are most *informative* for predicting the target. DTs can be interpreted as a tree or as a set of if-else rules which is a useful representation for human understanding. The most successful DT models like Classification and Regression Trees (CART) [5] and C4.5 [22] are greedy search algorithms. Finding DTs by optimising for say a fixed size, is NP-hard, with no polynomial-time approximation [9]. Modern algorithms have attempted this by enforcing constraints such as the independence of variables [10] or using all-purpose optimization toolboxes [2,4,27].

In [12] authors attempt the optimisation of the algorithm for model interpretability to derive decision lists. The reduced size of the rules opens up the option of interpreting the decisions in their entirety and not in the context of a specific input/output alone - a global explanation. The authors highlight the influence of complexity on the understandability of end-users. However, decision list algorithms still do not scale well for larger datasets. Optimal Sparse Decision Trees (OSDT) [8] and later improved with Generalized and Scalable Optimal Sparse Decision Trees (GOSDT) [14] algorithms produce optimal decision trees over a variety of objectives including F-score, AUC, and partial area under the ROC convex hull. GOSDT generates trees with a smaller number of nodes while maintaining accuracy on par with state-of-art models.

On explaining DTs for end-users, current studies have investigated local explanations using approaches such as counterfactuals [28], the integration of contextual information and identified narrative style textual explanations [17]. All these attempts to answer the *why* questions based on a few input features and specific to a particular input. Extending these insights to global explanations should help better understanding of the model by end-users and allow generalisation of the interpretations, driving actionability.

[1] https://qrisk.org/index.php.

3 Experiment Design

Our main research question is to determine what type of explanation are most relevant for non-expert end-users to be able to understand underlying risk model. We evaluate a local and global explanation by measuring user's perceived understanding and completeness. We also measure whether the user's mental model had changed after reading an explanation.

3.1 Dataset and Modeling

For the experiment, we used the Busselton dataset [11], which consists of 2874 patient records and information about whether they developed coronary heart disease (CHD) within ten years of the initial data collection. This study is similar to the data collected by NHS to develop QRISK3 [7]. Computing a risk score demands that we also explain the risk score, data used, probability measures of the scoring algorithm in addition to model prediction. We limit the scope of this study to only explaining the model prediction and use the CHD observation from the dataset as target variable for prediction. Using GOSDT [14] algorithm, we fit the data to obtain decision tress. GOSDT handles both categorical and continuous variables. While the optimum model may have multiple closeby splits for numeric values, such splits can reduce the readability of the tree. Hence we preprocess the data by converting most of the features into categorical variables. We follow the categories as mandated by National Health Service (NHS). The data is pre-processed as described in Appendix A, with 2669 records and 11 features.

The GOSDT algorithm generated a comprehensive decision tree for the dataset, comprising 19 leaf nodes at a depth of 5, achieving an accuracy of 90.9% (Fig. 4 in Appendix A). For the purpose of human evaluation and comparison of local and global explanations, it was necessary to have multiple DTs with comparable structures. Hence, we created subsets of the data by varying the ranges and combinations of *Age* and *Gender*. By working with reduced data points, the size of the constructed trees was significantly reduced. To ensure larger trees for evaluation purposes, we enforced a consistent depth of 4. Ultimately, we selected four trees for the evaluation task as shown in Table 1.

As mentioned in [20], a higher complexity of explanation rules in clinical setting leads to longer response times and decreased satisfaction with the explanations for end-user. The authors refer to unique variables within the rules as cognitive chunks, which contribute to complexity in understanding. In our experiment, global explanations naturally contain more cognitive chunks. To prevent bias in the results, we incorporated two levels of difficulty for each explanation type. The easy level consisted of trees with similar structures, both local and global, featuring 5 nodes and decision paths of equal length with an identical number of cognitive chunks. For ease of understanding, we henceforth refer to a particular combination of explanation type and difficulty level as a specific scenario, namely - local-easy, global-easy, local-hard, and global-hard. A local-SHAP explanation was generated utilizing the same tree as the local-easy scenario. We use kernel SHAP [15] to obtain feature importance for the local-easy tree for specific patient input. The SHAP explanation is treated as a baseline for evaluation.

The hard scenario for both explanation types, consist of larger trees of similar structures. The tree had 8 nodes for local-hard scenario and 9 nodes in case of global-hard scenario. For global explanations, the explanation presentation involves more cognitive chunks, potentially introducing bias by making the global-hard scenario challenging to comprehend. Nevertheless, we proceeded with evaluating this scenario in our experiment.

Another factor to consider when generating explanation is the possible contradiction between model explanation and general assumptions. For instance, a node *BMI* = *Normal* appearing in decision rules for low CHD risk is expected but not in those for high risk. Communicating this contradiction in explanation would be important in its understandability. We also include this in our experiment. Explanation scenarios categorized as hard involved contradictory explanations, which could prove more challenging for comprehension. We addressed these cases using semifactual [19] explanations, employing phrase *even-if*. We assess the impact of such risk narrations on understandability. Table 1 provides a summary of the four trees used for explanation generation.

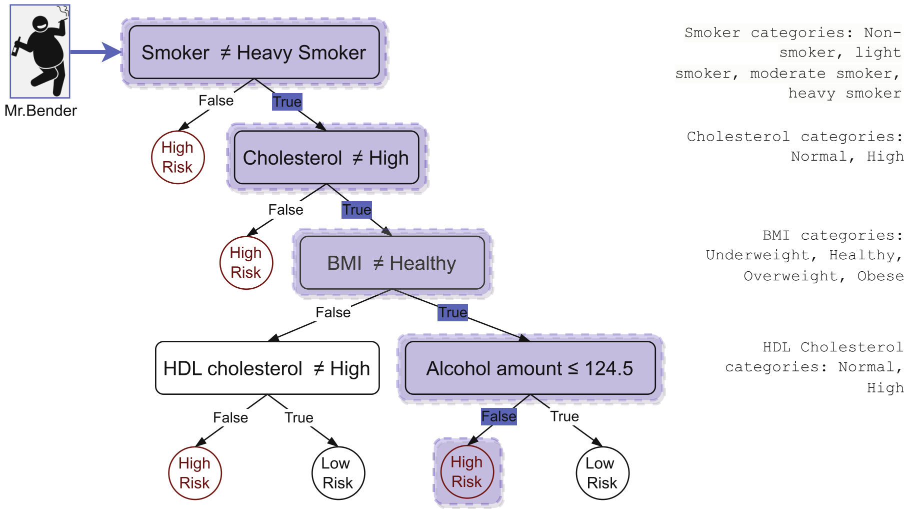
Table 1. Description of DTs and type of explanation generated.

Age	Gender	Leaf count	Accuracy	Explanation Type
70–79	Female	6	78.4	Local Easy
60–84	Female	6	82.5	Global Easy
60–70	Male	9	77.3	Local Hard
65–70	male	10	85.4	Global Hard

3.2 Generation of Explanation

For a given CHD prediction model and a corresponding patient input, the local explanation is a set of necessary conditions and predicted decisions of high/low risk. For the decision tree model in Fig. 2a, given particular patient info as input, the decision rule that is triggered to predict high risk is highlighted in blue. The path followed for the decision can be represented textually as shown in Fig. 2b. This is one possible representation. A more natural language expression of the rule is treated as a local explanation for the experiment. The language generation is rule-based. Details of the generation algorithm and an example of the evaluated explanation are given in Appendix B.

The global tree explanation is a list of all the decision rules of the tree. For a particular patient, a combination of the global explanation and the specific rule triggered corresponding to the given patient input is treated as the global


Mr.Bender
Smoker ≠ Heavy Smoker
False
True
High Risk
Cholesterol ≠ High
False
True
High Risk
BMI ≠ Healthy
False
True
HDL cholesterol ≠ High
Alcohol amount ≤ 124.5
False
True
High Risk
Low Risk
False
True
High Risk
Low Risk
Smoker categories: Non-smoker, light smoker, moderate smoker, heavy smoker
Cholesterol categories: Normal, High
BMI categories: Underweight, Healthy, Overweight, Obese
HDL Cholesterol categories: Normal, High

(a) The decision path followed along a given DT for a particular patient Mr Bender. The tree is learned from different categorical features of a patient dataset and the black square boxes represent decision nodes learned by the model. On the right, all the possible values of each feature are listed (except Alchohol amount which is numeric). This tree has 6 leaf nodes each a possible decision of *high* or *low risk*. For a given input corresponding to Mr Bender, the model predicts *high risk* following the decision path highlighted in Blue.

Mr.Bender has High risk of CHD since (BMI is not Healthy) and (daily alcohol consumption is greater than 124.5ml) even if is (not Heavy Smoker) and (Cholesterol is not High).

(b) A local explanation of the decision in 2a.

A patient has High risk of CHD if:

1. (Is Heavy Smoker).
2. (Cholesterol is High) even if (not Heavy Smoker).
3. (HDL cholesterol is High) even if is (not Heavy Smoker) and (Cholesterol is not High) and (BMI is Healthy).
4. (BMI is not Healthy) and (daily alcohol consumption is greater. than 124.5ml) even if (not Heavy Smoker) and (Cholesterol is not High).

Mr.Bender has High risk of CHD since he follows Rule 4.

(c) A global explanation of the DT and the decision in 2a.

Fig. 2. An example of local and global narrative explanation of a DT. Note that this is one way of generating a global tree explanation (Appendix B). Listing all the nodes or stating all possible categorical values of features are design choices that will affect understandability.

prediction explanation. Once again, this is a choice we make for this experiment. A list of all decision nodes similar to feature importance in SHAP could also be a possible global tree explanation. For the patient in Fig. 2a, the corresponding global explanation is shown in Fig. 2c. As the tree size becomes large, the number of rules and the number of features in each rule increase. This means the explanation size and the cognitive chunks in the explanation increase. The best way to frame natural language explanations, for these different cases, is a separate research problem that we do not address here. Further, we restrict the rules in global explanation to those corresponding to a single risk category - high risk. Since the particular case involves only two categories, this still provides coverage to possible predictions while keeping the explanation less verbose. The narration generation involves the same algorithm as in the case of local explanation.

In addition to the model accuracy, note that each leaf node has a probability and confidence associated with that particular decision. For a particular node, the probability is the ratio of training data points that fits the criteria of that node to the number of data points in its previous node. A low probability node denotes that, the particular decision was rare based on the training data. The statistical significance of this prediction denotes its confidence. Both these measures are used for generating decision narration. Appendix B shows examples of the usage. To express the probabilities, we use verbal mapping proposed by [26]. An additional usage of *possibly* is introduced to accommodate cases involving low confidence and high probability.

The SHAP explanation does not have associated confidence. We filter features with SHAP score greater that 0 and present them as bulleted points in descending order of importance.

3.3 Evaluation

For evaluation, a within-subject survey is conducted with participants recruited on Prolific platform. We conducted a pilot study among peers and the feedback was used to improve the readability of the explanations and assess the time taken for the tasks.

The survey involves 5 patient scenarios namely local-SHAP, local-easy, local-hard, global-easy and global-hard. Each scenario consists of 2 pages. On the first page, the participant is provided with information about a patient. This consists of their features: age, gender, height, weight, BMI, blood pressure, different cholesterol values, smoking, and drinking habit. They are asked to enter the assumptions on what patient features may contribute to the AI model's prediction. This captures the mental model of the participant regarding CHD. Appendix C shows examples of the pages used in the survey.

On the next page, participants are presented with the same patient, the risk of CHD (high or low) as predicted by the AI system along with an explanation. They are asked to enter feature importance once again based on their understanding of the explanation. They are also asked to rate the explanation on three parameters: completeness, understandability, and verboseness, using a 5-

Table 2. Evaluation criteria for comparison of different explanation types.

Measure	Definition
Completeness rating (CR)	User rating for the prompt: This explanation helps me completely understand why the AI system made the prediction
Understandability rating (UR)	User rating for the prompt: Based on the explanation I understand how the model would behave for another patient
Verboseness rating (VR)	User rating for the prompt: This explanation is long and uses more words than required
Change in mental model (CMM)	Difference in perceived feature importance before and after viewing model explanation
Error in Understanding (EU)	Difference between model feature importance and perceived feature importance after viewing explanation

level Likert scale. Text feedback on each explanation and overall feedback at the end of the survey is collected.

The evaluation of each explanation has 3 parameters from a Likert rating based on participant perceptions. In addition, based on the task of choosing feature importance we compute two additional parameters: change in mental model and correctness of understanding. Change in mental model is defined as the updation of perceived feature importance before and after explanation. Let $U = (u_1, u_2, ..., u_N)$ where $u_i \in \{0, 1\}, 1 \leq i \leq N$ be the selected feature importance before explanation where N is the total number of features. Let $V = (v_1, v_2, ..., v_N)$ where $v_i \in \{0, 1\}, 1 \leq i \leq N$ be the selected feature importance after explanation. *Change in mental model* is computed as

$$D_m = \frac{d(U, V)}{N}$$

where d is the *Hamming distance* between U and V.

For each explanation, based on the features that are shown in the narration, we also know the *correct* feature importance. In the case of SHAP, these are the features with a SHAP score greater than 0. For local explanations, these are the features in the decision path, and for global explanations, it is all the features in the tree. If the correct feature importance $C = (c_1, c_2, ..., c_N)$ where $c_i \in \{0, 1\}, 1 \leq i \leq N$, we compute the *error in understanding* w.r.t to the system mental model as

$$D_c = \frac{d(V, C)}{N}.$$

Since for each feature, the participant selects a yes/no for importance, these measures do not capture the relative importance among features. Table 2 summarises all the evaluation parameters.

4 Results and Discussion

Fifty participants were recruited from the Prolific platform for the experiment, ensuring a balanced gender profile. All participants were presented with five patient-explanation scenarios and were requested to evaluate each of them. The survey took an average of 26 min to complete, and participants received a compensation of £6 each, as per the minimum pay requirement. However, one participant was excluded from the analysis due to indications of low-effort responses, spending less than 1 min on multiple scenarios. The demographic details of the selected participants are summarized in Table 3. Based on the responses, we computed the evaluation parameters mentioned in the previous section. The Likert scale ratings for *Completeness*, *Understandability*, and *Verboseness* are assigned values from 0 to 1, 0 corresponding to 'Strongly Disagree' and 1 to 'Strongly Agree'. We also calculate, *Change in the mental model* and *Error in understanding* from the selection of feature importance. The calculated scores are also normalised to range from 0 to 1. The mean values across all participants are presented in Table 4.

Table 3. Demographic distribution of survey participants.

Feature	Category: Proportion
Age	18–30: 81.63%, 30–40: 16.33%, 40–65: 2.04%
Gender	Male: 51.02% , Female: 48.98%
First language	English: 38.8%, Others: 61.2%

Table 4. Evaluation parameters across all the scenarios. Maximum is highlighted in bold and minimum in italics. CR - Completeness rating, UR - Understandability rating, VR - Verboseness rating, CMM - Change in mental model, EU - Error in Understanding.

	Local SHAP	Local Easy	Local Hard	Global Easy	Global Hard
CR	0.64	0.69	*0.63*	0.68	**0.69**
UR	*0.66*	0.71	0.67	0.72	**0.74**
VR	*0.16*	0.26	0.23	**0.56**	0.52
CMM	**0.42**	*0.28*	0.38	0.34	0.35
EU	0.12	*0.07*	0.13	0.19	**0.30**

While local-easy scenario has the lowest error in understandability (EU), participants rated all the models comparably in terms of Understandability (UR) and Completeness (CR). The Change in the Mental Model (CMM) exhibited uniformity across all types of explanations, except for local-SHAP and local-easy. To assess the significance of these results, we performed the Wilcoxon test, for all combinations of explanation types. Since multiple comparisons are performed, we apply Bonferroni Correction on p-value and a threshold of 0.01 is chosen. In comparing local and global explanations, local-SHAP is excluded and the ratings for both levels of difficulty in each case are averaged. The results are shown in Table 5. The observations that hold for a stricter threshold of 0.001 are highlighted with $*$.

Table 5. Significance of difference between types of explanation. CR - Completeness rating, UR - Understandability rating, VR - Verboseness rating, CMM - Change in mental model, EU - Error in Understanding. The values which are significant (Bonferroni Corrected p-value threshold of 0.01) are highlighted in **bold**. P-value ≤ 0.001 are highlighted with *.

	CR	UR	VR	CMM	EU
Local vs Global	0.42	0.44	**0.00***	0.53	**0.00***
Local Easy vs Global Easy	0.84	0.85	**0.00***	0.05	**0.00***
Local Hard vs Global Hard	0.35	0.42	**0.00***	0.36	**0.00***
Local Easy vs Local Hard	0.38	0.24	0.76	**0.00***	**0.00***
Global Easy vs Global Hard	0.50	0.53	0.56	0.43	**0.00***
Local SHAP vs Local Hard	0.63	0.76	0.10	0.23	0.42
Local SHAP vs Local Easy	0.18	0.28	0.03	**0.00***	0.11
Local SHAP vs Global Hard	0.02	0.30	**0.00***	0.09	**0.00***
Local SHAP vs Global Easy	0.16	0.28	**0.00***	**0.01**	0.02

Global explanations resulted in a lower average understandability based on the feature selection (EU) and it was observed that harder scenarios resulted in higher errors for both local and global explanations. For each type of explanation, the patient features wrongly selected was investigated (Tables 11, 12). Incorrect feature selection related to *cholesterol* caused the majority of errors. Participants chose the wrong cholesterol-related feature, possibly due to a lack of attention or limited understanding of medical terminology. Improving the presentation of explanations and providing more contextual information could potentially address this issue. Importantly, when presented with semifactual explanations of hard scenarios both local and global explanations led to almost half or more participants excluding the corresponding feature. This clearly points to the ambiguity of such narration.

The error analysis does not explain the contradiction between the understandability ratings and the correctness of feature selection. Interestingly, a considerable number of participants expressed a preference for longer, global explanations, even if they did not fully comprehend them. Significant rating of global explanations as more verbose adds to this contradiction. To delve deeper into this phenomenon, participant clustering was performed based on the ratings and computed scores. Using the k-means algorithm, three distinct groups of participants were identified and manually validated. Figure 3 displays the average rating across different parameters for each group.

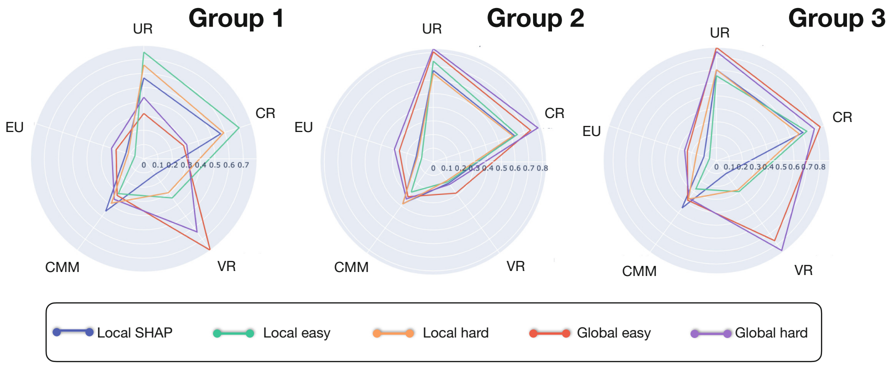


Fig. 3. Average rating for different explanation type across the participant groups

- Group 1: Strongly prefer and understand local explanations. The cluster consists of 11 participants who rate patient-specific local tree explanations highest on completeness and understandability.
- Group 2: Majority group that rates global explanation as most understandable: This cluster consist of 22 people who has the least significance in preference between global, local explanation or difference based on the difficulty level. They rate Global explanation highest on completeness and understandability
- Group 3: This cluster consist of 16 people who strongly prefer global explanations but critical about the narration. This cluster is more detail oriented and rates global explanations as more understandable and complete. This group was critical on the narration and presentation of explanation in the feedback form. The average error in feature selection for global explanation for this group, is lower than Group 2.

It is evident that within the clusters, the ratings on each parameters has significant preferential pattern between each type of explanation. Group 1, 3 has strong polarity on the preferences and their rating tend to *Strongly agree* or *Strongly disagree*. Both these Groups identify Global explanations as verbose. This shows that, in healthcare setting, the effectiveness of an explanation to an end-user, is very dependent on their individual preference.

4.1 Local vs. Global

While there is no significant difference between local and global explanations overall, strong differences emerge at the Group level. Group 1 rates the local explanation as complete, while both Groups 2, and 3 favour the global explanation for completeness. Similar preferences are observed in participants' perception of understandability within each group. When a stricter p-value threshold of 0.001 is applied, the significance of the difference in user rating for understandability and correctness holds only in Group 1. The results of the Wilcoxon test for combinations of explanation types within Groups are given in Appendix D.

- *The results indicate that certain people strongly prefer specific type of explanation. This preference does not necessarily translate to understandability.*
- *In all groups, a higher error in feature selection is observed for global explanations, mainly due to the semifactual explanation and wrongly interpreting features related to Cholesterol*

Among participants belonging to Group 2, the factors driving their preference for global explanations remain unclear. Demographics data examination (Table 6) offers no apparent patterns, leading us to propose the influence of unique cognitive styles within the groups. Further investigations are warranted to unveil the underlying reasons for these preferences and errors. While users may perceive explanations as understandable, it is vital to recognize that this perception may not necessarily translate into accurate decision-making. The lack of significant changes in mental models substantiate this, indicating the need for continued exploration to optimize explanation presentations for healthcare AI models.

Table 6. Demographic distribution of participants within each group. All the features are not available for all participants. Missing data are excluded in the counts.

	Group1	Group2	Group3
Number of participants	11	22	16
Male to female ratio	4:7	9:13	12:4
Count of full time employed	2	8	5
Student to non-student ratio	8:2	10:9	8:7
Number of native english speakers	4	11	4
Ethnicity, white to black ratio	9:2	11:10	11:3

4.2 Tree Explanation vs. SHAP

The overall ratings of SHAP explanations are comparable to those of local-hard explanations but lower than those of local-easy explanations generated from the same underlying decision tree. This suggests that the comprehensibility and interpretability of SHAP explanations are slightly lower than those of the local-easy explanations. However, this may be attributed to the presentation bias, as all participants were exposed to the SHAP explanation first. It is noted that the presentation style of SHAP explanations, using bulleted points, is generally considered less verbose even though it does not impact the error in understandability or perceived understandability and completeness. Hence the simpler readability of the SHAP explanation is not seen to have impacted its overall understandability.

4.3 Easy vs. Hard

The ratings provided by the participants on the Likert scale did not reveal any significant distinction between the explanation scenarios characterized as easy and hard. However, an examination of the impact of difficulty levels on the error in feature selection uncovered significant results. Hard scenarios, whether global or local explanations, exhibited significantly higher error rates, even within participant groups.

- *The explanation understanding is strongly dependent on the complexity of the feature interaction being explained.*

When participants encountered explanations that deviated from their preexisting notions of feature dependence, it introduced confusion, becoming a major contributor to error in hard scenarios. We observed that harder scenarios, on average, caused larger changes in the mental model of participants. However, this alone was insufficient to mitigate the observed errors. Furthermore, the consistent error patterns across different participant groups present an opportunity to enhance the current framework of narration and presentation of explanations, benefiting all participants.

5 Limitations and Future Work

The experiment provides evidence for the usefulness of global explanations in health informatics. Identifying cognitive styles that lead to particular explanation preferences and errors in comprehension, is pivotal to applying global explanations in real-life applications. The current experiment has been carried out on a small dataset. Evaluating these findings on a larger data set with more data points and larger features will be undertaken in future studies. We recognise that regression models are commonly used in risk prediction. Expanding the scope of the narrative global explanation within the context of regression and assessing its comparative utility against the local explanation will enable the integration of our findings into established risk predictive tools.

Further, the evaluation in this study was crowdsourced and hence the participants are not representative of real-life patients. Most of the participants fall in the age category that does not have a risk of heart disease as predicted by the model. This may have biased their rating. We aim to rectify this by conducting the evaluation on a representative patient population, which would also require addressing ethical concerns.

The current study has not focussed on generating effective global explanations. The use of semifactuals has not addressed the mismatch with the user's mental models. Further, the presentation of Explanation features is seen to have introduced errors. Effective communication and presentation techniques would be vital in reducing errors. Though we have used a linguistic representation of probability and confidence, the evaluations in this regard remain undone. For risk communication at scale, this is a crucial component. Further research is warranted to delve deeper into these aspects and refine the design and implementation of explanation systems.

Acknowledgement. We would like to thank Dr. Sameen Maruf and Prof. Ingrid Zukerman for generously sharing their expertise in utilizing the dataset, continuous support and valuable feedback in designing the experiment. We thank Nikolay Babakov and Prof. Alberto José Bugarín Diz for their feedback throughout the development of this research. We also thank the anonymous reviewers for their feedback which has significantly improved this work. A. Sivaprasad is ESR in the NL4XAI project which has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 860621.

A Construction and Selection of DT

B Generating Explanation Narration

Steps in generating narration (Figs. 5, 6 and 7):

1. Filter the rules corresponding to high risk leaves.
2. Sort the decision rules in order of their leaf node confidence and insert verbal mapping of relative probability.
3. Reorder the features and place contradictory features at the end preceded by even-if.
4. combine the features with *and*
5. Add header with *age*, *gender*

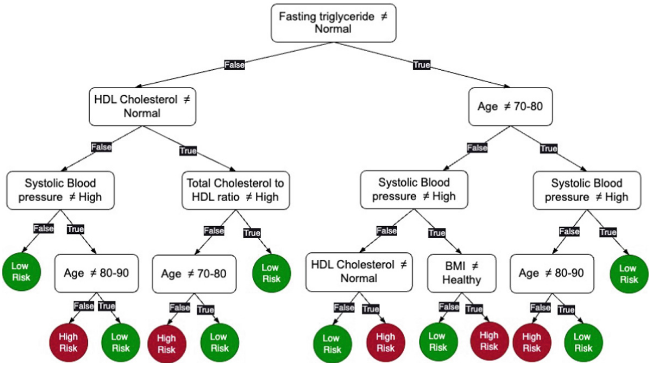


Fig. 4. Depth 5 Decision tree generated on 2134 datapoints. Training accuracy = 90.9%, test accuracy on 534 records = 85%.

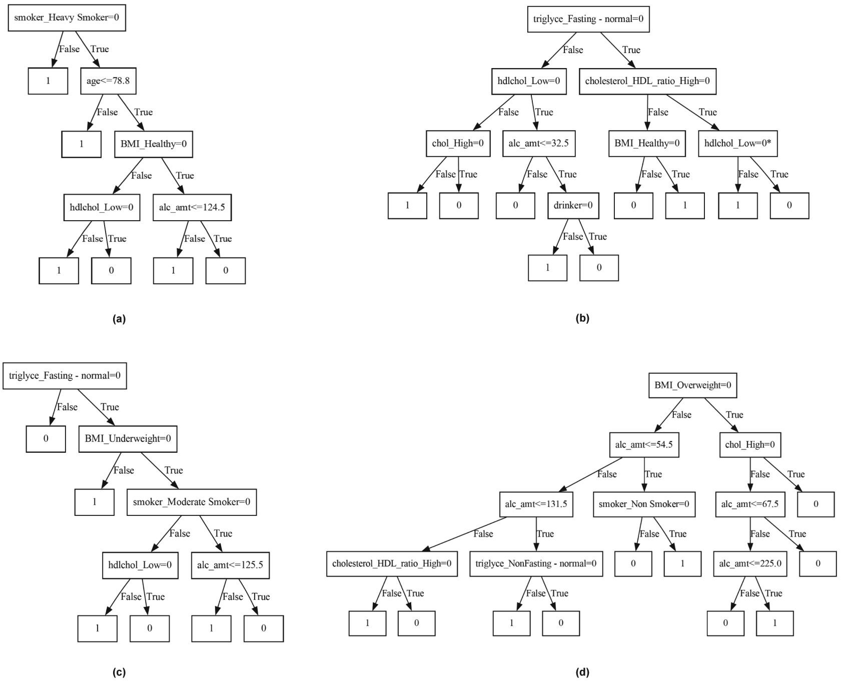


Fig. 5. DTs for different scenarios. (a) Local easy scenario: Decision tree generated on 116 data points. Training accuracy = 78.4%, (b) Local Hard scenario: Decision tree generated on 163 data points. Training accuracy = 77.3%, (c) Global easy scenario: Decision tree generated on 382 data points. Training accuracy = 82.5% (d) Global Hard scenario: Decision tree generated on 108 data points. Training accuracy = 85.4%.

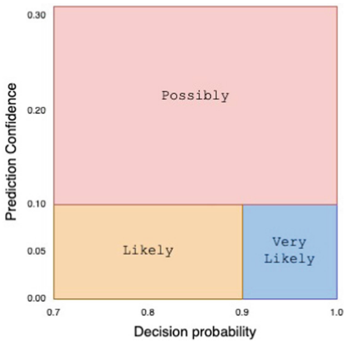
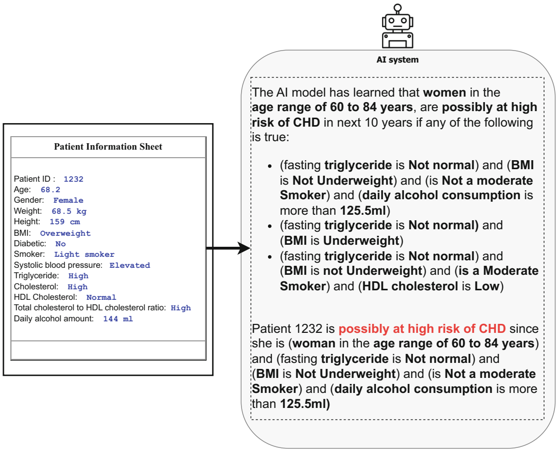


Fig. 6. Verbal mapping of relative probabilities.



Patient Information Sheet

Patient ID : 1232
Age: 68.2
Gender: Female
Weight: 68.5 kg
Height: 159 cm
BMI: Overweight
Diabetic: No
Smoker: Light smoker
Systolic blood pressure: Elevated
Triglyceride: High
Cholesterol: High
HDL Cholesterol: Normal
Total cholesterol to HDL cholesterol ratio: High
Daily alcohol amount: 144 ml

AI system

The AI model has learned that **women** in the **age range of 60 to 84 years**, are **possibly at high risk of CHD** in next 10 years if any of the following is true:

- (fasting **triglyceride** is **Not normal**) and (**BMI** is **Not Underweight**) and (is **Not a moderate Smoker**) and (**daily alcohol consumption** is more than **125.5ml**)
- (fasting **triglyceride** is **Not normal**) and (**BMI** is **Underweight**)
- (fasting **triglyceride** is **Not normal**) and (**BMI is not Underweight**) and (**is a Moderate Smoker**) and (**HDL cholesterol** is **Low**)

Patient 1232 is **possibly at high risk of CHD** since she is (**woman** in the **age range of 60 to 84 years**) and (fasting **triglyceride** is **Not normal**) and (**BMI** is **Not Underweight**) and (is **Not a moderate Smoker**) and (**daily alcohol consumption** is more than **125.5ml)**

Fig. 7. An example for generated global explanation. This model corresponds to Global-easy scenario.

Table 7. Category definitions for Data preprocessing.

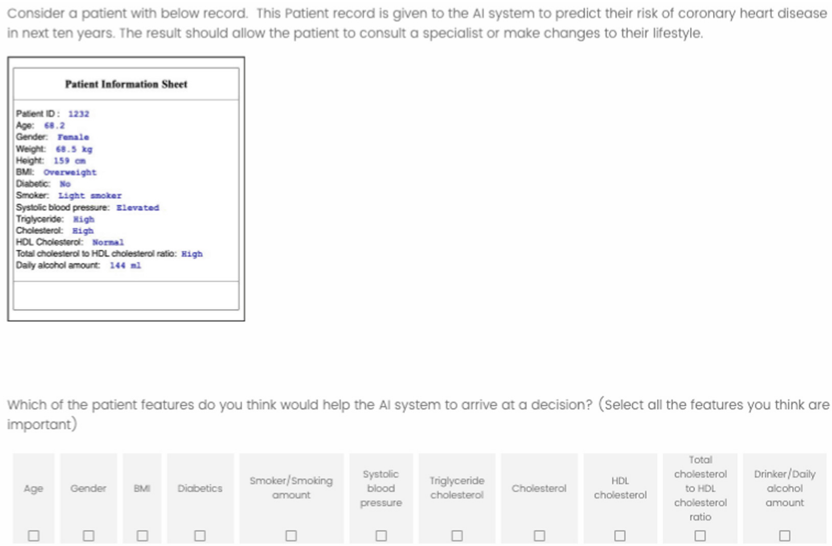
Feature	Categories
Smoking	Non-smoker, light smoker: (less than 10), moderate smoker - (10 to 19)/day, heavy smoker- (20 or over)/day)
BMI	Underweight (less than 18.5), Healthy - (18.5 to 24.9), Overweight - (25 to 29.9), Obese - (30 or over)
Cholesterol	Normal: ≤ 5, High: above 5
Cholestrol HDL ratio	Normal: ≤ 6, High: above 6
Triglycerides	Fasting - normal (0 to 1.7), Non Fasting - normal (1.7 to 2.3), High (2.3 to 10)
HDL	Normal: $\leq$ 1, High: above 1
Systolic Blood Pressure	Low: (0 to 90), Normal: (90 to 120), Elevated: (120 to 140), High: (140 to 250)
Diastolic Blood pressure	Low: (0 to 60), Normal: (60 to 80), Elevated: (80 to 90), High: (90 to 150)

C User Survey on Prolific

For each scenario, a participant first see the patient information as shown in Fig. 8. The participant is asked to pick all the features they think might be influential in predicting the patient's risk of CHD. This captures the participants mental model regarding CHD prediction before viewing any explanation (Table 7).

In the next page, a participant is shown the explanation followed by questions to rate the explanation. The are asked to redo the task of picking all the features they think were influential in predicting the patient's risk of CHD as shown in Fig. 9. This captures the participant's understanding of AI's mental model. This is followed by questions to get the users rating based on a 5 point Likert scale. The questions correspond to 3 parameters being measured:

1. Completeness: This explanation helps me completely understand why the AI system made the prediction
2. Understandability: Based on the explanation, I understand how the model would behave for another patient
3. Verboseness: This explanation is long and uses more words than require



Consider a patient with below record. This Patient record is given to the AI system to predict their risk of coronary heart disease in next ten years. The result should allow the patient to consult a specialist or make changes to their lifestyle.

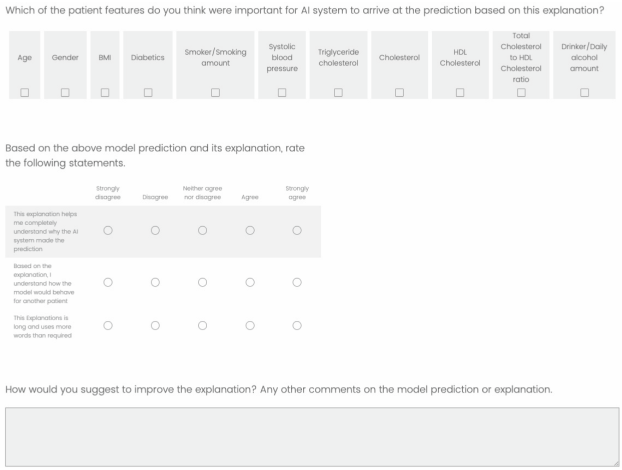
Patient Information Sheet

Patient ID : 1232
Age: 68.2
Gender: Female
Weight: 68.5 kg
Height: 159 cm
BMI: Overweight
Diabetic: No
Smoker: Light smoker
Systolic blood pressure: Elevated
Triglyceride: High
Cholesterol: High
HDL Cholesterol: Normal
Total cholesterol to HDL cholesterol ratio: High
Daily alcohol amount: 144 ml

Which of the patient features do you think would help the AI system to arrive at a decision? (Select all the features you think are important)

Age	Gender	BMI	Diabetics	Smoker/Smoking amount	Systolic blood pressure	Triglyceride cholesterol	Cholesterol	HDL cholesterol	Total cholesterol to HDL cholesterol ratio	Drinker/Daily alcohol amount
□	□	□	□	□	□	□	□	□	□	□

Fig. 8. First page of a scenario shown to participants with a patient info. They question captures the participant's mental model of CHD prediction before viewing explanation.



Which of the patient features do you think were important for AI system to arrive at the prediction based on this explanation?

Age	Gender	BMI	Diabetics	Smoker/Smoking amount	Systolic blood pressure	Triglyceride cholesterol	Cholesterol	HDL Cholesterol	Total Cholesterol to HDL Cholesterol ratio	Drinker/Daily alcohol amount
□	□	□	□	□	□	□	□	□	□	□

Based on the above model prediction and its explanation, rate the following statements.

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
This explanation helps me completely understand why the AI system made the prediction	○	○	○	○	○
Based on the explanation, I understand how the model would behave for another patient	○	○	○	○	○
This Explanations is long and uses more words than required	○	○	○	○	○

How would you suggest to improve the explanation? Any other comments on the model prediction or explanation.

Fig. 9. The first question evaluates participant's understanding of the explanation. The Remaining questions capturing their feedback on explanation.

D Comparison of Local and Global Explanation Ratings

Results of Wilcoxon test, for combinations of explanation types within participant Groups. After Bonferroni Correction, p-values less than 0.01 are chosen as significant (Tables 8, 9 and 10).

Table 8. Significance of difference between different types of explanation for Group 1 rounded to 2 decimal places. Significant p-value are in **bold**. P-value ≤ 0.001 are highlighted with *.

	CR	UR	VR	CMM	EU
Local vs Global	**0.00***	**0.00***	**0.00***	0.81	**0.00**
Local easy vs Global easy	**0.00**	**0.01**	**0.00**	0.93	**0.00**
Local Hard vs Global Hard	0.05	0.05	0.03	0.62	0.07
Local easy vs Local Hard	0.20	0.21	0.50	0.04	0.19
Global easy vs Global Hard	0.34	0.65	0.16	0.56	0.66
Local SHAP vs Local Hard	0.53	0.79	0.04	0.29	0.79
Local SHAP vs Local easy	0.04	0.34	**0.01**	**0.01**	0.18
Local SHAP vs Global Hard	0.21	0.03	**0.01**	0.21	0.10
Local SHAP vs Global easy	0.03	0.07	**0.00**	0.02	0.35

Table 9. Significance of difference between different types of explanation for Group 2 rounded to 2 decimal places. Significant p-value are in **bold**. P-value ≤ 0.001 are highlighted with *.

	CR	UR	VR	CMM	EU
Local vs Global	0.02	**0.01**	0.13	0.89	**0.00***
Local easy vs Global easy	0.22	0.08	0.10	0.38	**0.00***
Local Hard vs Global Hard	0.06	0.08	0.65	0.29	**0.00***
Local easy vs Local Hard	0.40	0.60	0.92	0.02	0.19
Global easy vs Global Hard	0.53	0.24	0.21	0.42	0.35
Local SHAP vs Local Hard	0.84	0.99	0.51	0.81	0.92
Local SHAP vs Local easy	0.27	0.71	0.97	0.05	0.58
Local SHAP vs Global Hard	**0.00**	0.02	0.80	0.71	**0.01**
Local SHAP vs Global easy	0.03	0.07	0.10	0.20	0.02

Table 10. Significance of difference between different types of explanation for Group 3 rounded to 2 decimal places. Significant p-value are in **bold**. P-value $\leq$ 0.001 are highlighted with *.

	CR	UR	VR	CMM	EU
Local vs Global	**0.01**	0.05	**0.00***	0.17	**0.00***
Local easy vs Global Hard	0.02	0.10	**0.00**	0.03	**0.00***
Local Hard vs Global Hard	0.27	0.27	**0.00***	0.94	**0.01**
Local easy vs Local Hard	0.60	0.62	0.77	0.14	**0.00**
Global easy vs Global Hard	0.66	0.26	0.14	0.96	0.40
Local SHAP vs Local Hard	0.56	0.62	0.13	0.25	0.06
Local SHAP vs Local easy	0.60	0.60	0.07	**0.01**	0.29
Local SHAP vs Global Hard	0.05	0.23	**0.00***	0.15	**0.00***
Local SHAP vs Global easy	0.03	0.05	**0.00***	0.23	**0.01**

Table 11. Error in selecting patient feature after explanation. Type I error (False Positive) - Wrong selection overall.

	Local SHAP	Local easy	Local hard	Global easy	Global hard
Age					
Gender	3				
BMI			2		
Diabetics	5	2		1	
Cholesterol	5	2	2	8	
HDL cholesterol					15
Triglyceride cholesterol	1				
Total cholesterol to HDL cholesterol ratio	2	1	1	6	
Systolic blood pressure	5	1		2	5
Smoking/Smoking amount			2		
Dinker/Drinking amount					2

Table 12. Error in selecting patient feature after explanation. Type II error (False Negative) - Missing correct feature.

	Local SHAP	Local easy	Local hard	Global easy	Global hard
Age	6	6	1	8	9
Gender		8	13	22	23
BMI	14	1		1	3
Diabetics					
Cholesterol					31
HDL cholesterol	10		23	37	
Triglyceride cholesterol			20	4	35
Total cholesterol to HDL cholesterol ratio					11
Systolic blood pressure					
Smoking/Smoking amount	4	17		10	27
Dinker/Drinking amount	12		9	3	

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Bertsimas, D., Dunn, J.: Optimal classification trees. Mach. Learn. **106**(7), 1039–1082 (2017). https://doi.org/10.1007/s10994-017-5633-9
3. Biran, O., Cotton, C.V.: Explanation and justification in machine learning: a survey. In: IJCAI-17 Workshop on Explainable AI (XAI), vol. 8 (2017)
4. Blanquero, R., Carrizosa, E., Molero-Río, C., Morales, D.R.: Optimal randomized classification trees. Comput. Oper. Res. **132**, 105281 (2021). https://doi.org/10.1016/j.cor.2021.105281
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and regression trees (1984)
6. Glik, D.C.: Risk communication for public health emergencies. Annu. Rev. Publ. Health **28**(1), 33–54 (2007). https://doi.org/10.1146/annurev.publhealth.28.021406.144123, pMID: 17222081
7. Hippisley-Cox, J., Coupland, C., Brindle, P.: Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ **357** (2017). https://doi.org/10.1136/bmj.j2099
8. Hu, X., Rudin, C., Seltzer, M.: Optimal sparse decision trees. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 32. Curran Associates, Inc. (2019)
9. Hyafil, L., Rivest, R.L.: Constructing optimal binary decision trees is np-complete. Inf. Process. Lett. **5**(1), 15–17 (1976)
10. Klivans, A.R., Servedio, R.A.: Toward attribute efficient learning of decision lists and parities. In: Shawe-Taylor, J., Singer, Y. (eds.) COLT 2004. LNCS (LNAI), vol. 3120, pp. 224–238. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-27819-1_16
11. Knuiman, M.W., Vu, H.T., Bartholomew, H.C.: Multivariate risk estimation for coronary heart disease: the Busselton health study. Aust. N. Z. J. Publ. Health **22**(7), 747–753 (1998)
12. Lakkaraju, H., Bach, S.H., Leskovec, J.: Interpretable decision sets: a joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 1675–1684. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2939672.2939874
13. Letham, B., Rudin, C., McCormick, T., Madigan, D.: Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model. Ann. Appl. Stat. **9**, 1350–1371 (2015). https://doi.org/10.1214/15-AOAS848
14. Lin, J., Zhong, C., Hu, D., Rudin, C., Seltzer, M.: Generalized and scalable optimal sparse decision trees. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning, vol. 119, pp. 6150–6160. PMLR, 13–18 July 2020
15. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17, pp. 4768–4777. Curran Associates Inc., Red Hook, NY, USA (2017)

16. Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inform. **113**, 103655 (2021). https://doi.org/10.1016/j.jbi.2020.103655
17. Maruf, S., Zukerman, I., Reiter, E., Haffari, G.: Influence of context on users' views about explanations for decision-tree predictions. Comput. Speech Lang. **81**, 101483 (2023). https://doi.org/10.1016/j.csl.2023.101483
18. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
19. Moreno-Ríos, S., García-Madruga, J.A., Byrne, R.M.: Inferences from semifactual 'even if' conditionals. Acta Physiol. (OXF) **128**(2), 197–209 (2008). https://doi.org/10.1016/j.actpsy.2007.12.008
20. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. CoRR abs/1802.00682 (2018)
21. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect, 1st edn. Basic Books Inc., New York (2018)
22. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993)
23. Reiter, E.: Natural language generation challenges for explainable AI. In: Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artif. Intell. (NL4XAI 2019), pp. 3–7. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/W19-8402
24. Ribeiro, M., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 97–101. Association for Computational Linguistics, San Diego, California (2016). https://doi.org/10.18653/v1/N16-3020
25. Sendak, M.P., Gao, M., Brajer, N., Balu, S.: Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit. Med. **3** (2020)
26. Spiegelhalter, D.: Risk and uncertainty communication. Annu. Rev. Stat. Appl. **4**(1), 31–60 (2017). https://doi.org/10.1146/annurev-statistics-010814-020148
27. Verwer, S., Zhang, Y.: Learning optimal classification trees using a binary linear program formulation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 1625–1632 (2019)
28. Wachter, S., Mittelstadt, B.D., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. CoRR abs/1711.00399 (2017)