An Interactive XAI Interface with Application in Healthcare for Non-experts

Jingyu Hu[1], Yizhu Liang[1,2], Weiyu Zhao[1], Kevin McAreavey[1], and Weiru Liu[1]

[1] University of Bristol, UK
[2] Bank of China Fintech, China
jingyu-hu@outlook.com,
{dn21257, jm21920, kevin.mcareavey, weiru.liu}@bristol.ac.uk

Abstract. Explainable artificial intelligence (XAI) has gained increasing attention in the medical field, where understanding the reasons for predictions is crucial. In this paper we introduce an interactive and dynamic visual interface providing global, local and counterfactual explanations to end-users, with a use case in healthcare. The dataset used in the study is about predicting an individual's coronary heart disease (CHD) within 10 years using the decision tree classification method. We evaluated our XAI system with 200 participants. Our results show that the participants reported an overall good assessment of the user interface, with non-expert users showing a higher satisfaction than users who have some degree of knoweldge of AI.

Keywords: XAI · non-expert users · interactive XAI system · global, local, counterfactual explanation.

1 Introduction

1.1 Background

Artificial intelligence (AI), the intelligent technology of machines [30], has made significant progress in performing tasks that traditionally require human intelligence. Recent advances in both hardware and high-performance computing have enabled the development of increasingly complex AI models that achieve high accuracy by continuously turning their parameters.

However, the increasing complexity and the lack of transparency of AI models, especially black-box models, make it difficult to convey security and trustworthiness to users in how and why decisions are made in different applications [9]. The lack of transparency has led to ethical concerns in various fields, such as medical diagnosis and legal judgment. The ability to explain why a certain decision was made, became a vital property of AI systems. Derived from the emerging demand for explaining the deployed AI systems, the topic of Explainable Artificial Intelligence (XAI) emerged recently, and has since became an active research area. XAI aims to make AI more understandable by providing details and reasons for its decisions and actions [3].

The concepts of explainability and interpretability: In the community of XAI, the concepts of explainability and interpretability are equated in some cases but there does exist a subtle difference between them. According to the paper by Biran and Cotton [5], systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation. In general, the process of interpretation for a model tends to be revealing the working structure and the rationale, while explanation mostly focuses on providing post-hoc explanations for existing machine learning models. The AWS Whitepaper [3] gives a brief summary. Interpretability focuses more on the inner mechanics of models, which are about how and why the predictions are generated, raising the question *How does the model work?*. Explainability is the ability to explain the model's behaviours in human terms, which can usually be achieved by model-agnostic methods, raising *What else can the model tell me?*. There exists a tradeoff between interpretability and model performance given common AI/ML models.

1.2 The methods of XAI

While there has been an explosive growth of XAI methods, they have two common broad aims: transparency and post-hoc interpretation [23]. Transparency refers to how a model works intrinsically, while post-hoc interpretation concerns how a model behaves after the model training.

Based on a comprehensive and holistic analysis of previous surveys [1], XAI methods are organized into three categories: (1). Complexity-related methods (2). Scope-related methods (3). Model-related methods.

Table 1. XAI methods categories

Complexity Related Methods		
Scope Related Methods	Global Methods	
	Local Methods	
Model Related Methods	Model-specific Methods	
	Model-agnostic Methods	Visualization: Surrogate models, Partial Dependence Plot,etc
		Knowledge Extraction: Rule extraction, etc
		Influence Methods: Sensitivity analysis, Feature importance,etc
		Example-based Explanation: Counterfactual explanations, etc

In this study, we place emphasis on the following explanation approaches:

– Global explanation: global explanations focus on the overall logic of a model and the entire decision-making process that lead to all the different outcomes. This class of methods are applied when the macro-level decisions are crucial. A general strategy is to display the

[3] https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html

decision-making process by symbolic and graphical representations. For example, [10] introduced the algorithm of TREPAN, to generate symbolic representations for given neural networks and extract decision-tree structures. Partial Dependence plot (PDP) is a model-agnostic explanation method [12]. PDP displays the marginal effect of attributes to the output of models. Attributes can have either a linear or a more complex effect on the predicted outcome. The limitation of PDP derives from its assumption of independence that the attributes for which partial dependence are computed are not correlated with each other. Also, it is not able to describe the heterogeneous effects of attributes. The method of Individual Conditional Expectation (ICE) extends PDP. ICE plots reveal interactions and individual differences by disaggregating the partial dependence function, which enables a deeper understanding of the level of individual observations [1].

- Local explanation: local explanations aim at explaining why a particular decision was made. [29] presented the algorithm of LIME (Local Interpretable Model-agnostic Explanations), which is a model-agnostic method explaining the local decisions of any interpretable classifier or regressor. It can also approximate black-box models in a local neighborhood of any prediction. Given the goal of exploring the reason why the black-box model made a certain decision, LIME attempts to find out how the outcomes change when variations were added to the input data [24]. It feeds the black-box models by perturbed samples and generates a new dataset of perturbed samples plus the outcomes of models. Based on the new dataset, an interpretable model will be built up, which is weighted by the proximity of perturbed samples to the original [24]. LIME calculates and outputs how much each attribute contributes to the predication of a single sample. Another similar algorithm, Shapley Additive Explanations (SHAP), proposed by [19] in 2017, is a method from cooperative game theory, which assumes that each attribute value of the sample is a player in a game where the prediction outcome is the payout [24]. For each data sample, the algorithm computes the SHAP value of each attribute showing how much effect each attribute has on the prediction.
- Counterfactual explanation: counterfactual explanations describe the minimum alterations to the input data that are needed to obtain a different decision. Counterfactual methods do not touch the overall logic of the model, but focus on explaining individual predictions. Counterfactual explanations are useful when addressing questions such as "why the outcome is P rather than Q?". Therefore, counterfactual explanations can be understood as aiming to find data samples that can produce Q as the outcome while revealing which attributes' values will need to be changed (from the original data sample) in order to achieve this. In this sense, counterfactual expla-

nation approaches can be extended to provide contrastive explanations for non-classification problems.

XAI tools and applications: XAI methods have been applied to develop interpretive systems. In many fields, smart systems that incorporate domain knowledge and XAI are mostly used to assist experts. For example, Clinical Decision Support Systems (CDSSs) enhancing communication efficiency and assisting in the diagnosis by physicians [31]. Apart from those systems designed for professional and practitioners, there is a strong need to provide explanations to non-experts to facilitate the adoption and gain the public trust of AI in the wider society. Several interactive interfaces have been developed to address this need, including InterpretML [28], AIX360 [4], which both offer global and local explanations. [11] conducted a user study on eXplanation through Plan Properties (XPP) tools, and the evaluations indicate that these explanations enable users to find better trade-offs.

1.3 What constitutes a good explanation?

Though there are a great many ways to provide explanations, what constitutes a good explanation is still an issue requires considering. Recent works on XAI, which focused on simplified models that approximate the true criteria to make decisions, can lead to a gap of expectations between AI/ML and the fields of philosophy [23]. Given the questions such as "Are the explanations useful?", "Is the model understandable?", or "Is the decision-making sensible", people with different backgrounds may have opposite opinions. A good explanation in the view of a machine learning specialist may be unconvincing to the context of philosophy, sociology and cognitive sciences.

[22] provided a comprehensive review of social sciences on human explanation and discussed if and how these works can be applied to XAI. According to [22], humans have certain biases in their cognitive processes, which means they generate, select and evaluate explanations in a biased manner. When explaining a phenomenon, people are more likely to bias explanations towards inherent attributes, rather than extrinsic attributes. That bias towards inherence is thought to derive from prior knowledge, cognitive ability and so on [22]. As human explanations are selected, an explanation provider may not provide complete causes of an event, but can still convey useful information by emphasizing the key attributes or evidence in explanation based on their relevance to the recipients' interests. Explanations are social activities involving an interaction between explainers and explainees [23]. Explanations of AI/ML models can be conceived as generated by an iterative process, selected and evaluated based on presuppositions and beliefs [23].

[8] introduced their online experiment where participants use different interfaces to get explanations of an algorithm for making decisions on university admissions. By measuring users' understanding of the algorithm, it is found that interactive explanations are more effective than static explanations while "white-box" explanations are more effective than the "black-box". Those conclusions

can be conceived as the design principle of user interfaces, which enable users to explore the system's behaviours freely through interactive explanations and "white-box" (defined as the visualization of the inner working of the system in the paper) [8]. [6] conducted a controlled user study using 4 different systems to investigate if contextualizing and allowing the exploration of explanations based on local attribute importance could improve users' satisfaction. The results of analysis of variance demonstrated that by providing users with missing contextual information (ML knowledge, domain knowledge, external/real-life knowledge), and providing interactive attributes to test their hypotheses (interactive display and example-based explanations), the objective understanding scores of users are increased.

Therefore, those related works provide strong motivations for a user study with an interactive contextualized interface. Interacting with the interface, users engage in the communication through dialogue, textual description and graphical presentation which leads to their own understanding of the model.

2 Preliminaries

2.1 Global and Local Explanation for Decision Trees

Global Explanation: Global explanation is to explain how a model makes decisions by considering all the attributes. Decision tree [27] is a tree-like algorithm that recursively splits the data into smaller subsets and uses the tree leaves to represent the final classification result. In some decision tree algorithms, the maximum depth of the tree can be specified. In our experiments, we set the maximum tree depth to 6. One of the main criterion for splitting a node (D) in a tree into sub-branches (D1, D2, ...) is the GINI index, which calculates the effectiveness of a split based on an attribute at node D.

A commonly used formula for the GINI index is Equation 1, where $|D|$ (or $|D_i|$) represents the cardinality of set D (or D_i)

$$\text{GiniIndex}(D) = \sum_{i=1}^{n} \frac{D_i}{D} \text{Gini}\,(D_i) \tag{1}$$

Equation 2 is the definition of GINI, where c_j is the number of records in D_i with class label as j (for total of n class labels)

$$\text{Gini}(D_i) = 1 - \sum_{i=1}^{n} (\frac{c_j}{D_i})^2 \tag{2}$$

Figure 1, referenced from xoriant[4], presents an example of decision tree algorithm. The decision tree considers age, eating habits, and exercise preferences to determine whether a person is fit. The first layer is whether a person's age is less than 30. The second layer decides whether the person eats a lot of pizza or exercises in the morning. The leaf nodes are the final judgment result (fit/unfit).

[4] https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm

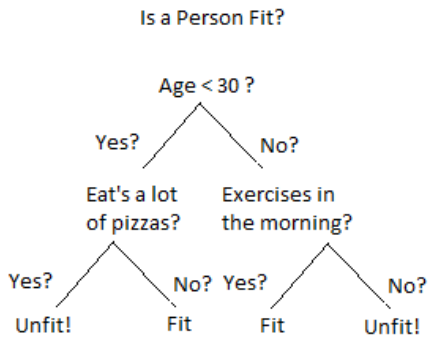


Fig. 1. A simple decision tree determining if a person is fit

Local Explanation: Local explanation is an explanation of how a model predicts a decision for a specific record. In the decision tree model, the specific decision path from the root to the leaf corresponding to that specific record can be regarded as a local explanation. For example, we can draw following local explanations about whether a person is fit or not by following each branch from root to node in the decision tree of Figure 1.

Example 1: if a person's age<30 AND person eats lots of pizzas THEN the person is unfit.

Example 2: if a person's age>30 AND person exercises in the morning THEN the person is fit.

2.2 Counterfactual Explanations

Counterfactual explanations aim to explain why a model predicted one result P instead of another result Q. Some early works in AI are closely related to counterfactuals, such as [14], [32]. However, the explanations provided by these expert-focused or rule-based systems do not offer insight into the internal logic of classifiers. Later counterfactual explanations took an end-to-end integrated approach. For data-driven classification[21], a heuristic method was proposed to explain classified documents. Meanwhile, adversarial perturbations[13], such as Deepfool attacks[25], have been studied for generating counterfactual explanations for deep neural networks. To overcome challenges in interpretability and accountability, researchers like [33] explored unconditional counterfactual explanations of automated decisions. Diverse Counterfactual Explanations(DiCE)[26] extends the work of [33] and provides a method that can be applied to any differentiable machine learning classifier.

Generate Counterfactual Records by DiCE: Equation 3 presents the original counterfactual explanation framework proposed by [33]. Given F as the

predictive ML model, it generates a counterfactual record c that has a different predicted outcome than that for the original record x by minimizing the loss function *yloss*.

$$\boldsymbol{c} = \arg\min_{\boldsymbol{c}} yloss(F(\boldsymbol{c}), F(\boldsymbol{x})) + Dis(\boldsymbol{x}, \boldsymbol{c}) \tag{3}$$

Where $Dis(x, c)$ is the distance between x and c. $Dis(x, c)$ keeps the counterfactual close to the original record and can be achieved with distance measures like Euclidean, Cosine and Manhattan distance.

DiCE introduces diversity and proximity constraints and optimises the above equation by presenting Equation 4, where $\lambda 1$ and $\lambda 2$ are hyperparameters used for balancing the weights of three parts in the equation.

Proximity and *Diverse* are defined as Equation 5 and Equation 6.

$$C(x) = \underset{c_1 \ldots, c_k}{\operatorname{argmin}} \frac{1}{k} \sum_{j=1}^{k} \text{yloss } (F(c_i), F(x)) + \lambda_1 \cdot Proximity + \lambda_2 \cdot Diverse \tag{4}$$

Proximity is quantified as the (negative) distance between the attributes of the original input and the generated record. In DiCE, the proximity of a set of counterfactual records is defined as their average proximity.

$$Proximity = -\frac{1}{k} \sum_{i=1}^{k} Dis(c_i, x) \tag{5}$$

Equation 6 says that the diversity of a set of counterfactuals $\{c_1, c_2, ...c_k\}$ is defined as $det(K)$, where the elements of matrix K equal to $K_{i,j}$. $det(K)$ is the determinant of K. It shows diversity constraints in subset selection problems implemented by Determinantal Point Processes (DPP)[17]. DPP-based diversity facilitates the selection of subsets containing more diverse elements, and results in higher probabilities for these subsets.

$$Diverse = \det(K), K_{i,j} = \frac{1}{1 + Dis(c_i, c_j)} \tag{6}$$

3 XAI System Design and Implementation

3.1 Architecture Design

This paper therefore reports the findings of Interactive Graphical User Interfaces (GUIs) of explainable Artificial Intelligent systems oriented to non-expert users, and to discuss how to improve the quality of explanations by means of user study. Given the dataset of cardiovascular, users are asked to explore a classification problem based on the model of decision tree. The interface provides explanations generated by global, local and counterfactual methods for users, helping them comprehend how and why the decision tree makes predictions. Based on

the existing methodologies of XAI and motivated by works that apply cognitive sciences, the design of interface takes advantage of the achievements from previous user studies, conforming to the principles of "interactive", "selective" and "contextualization".

Figure 2 displays the overall structure of the design. The web application is based on the Flask[15] framework. Bootstrap and LayUI templates are used to set the CSS styles. Tree visualizations are generated using Echarts[18] in Javascript. Data interactions are created using Jinja and Ajax requests. The front-end web pages and back-end implementations are packaged as Docker containers and deployed to a cloud server (Ubuntu 20). Nginx acts as a reverse proxy between users and cloud server, forwarding requests to the backend server application and returning its response to the user.

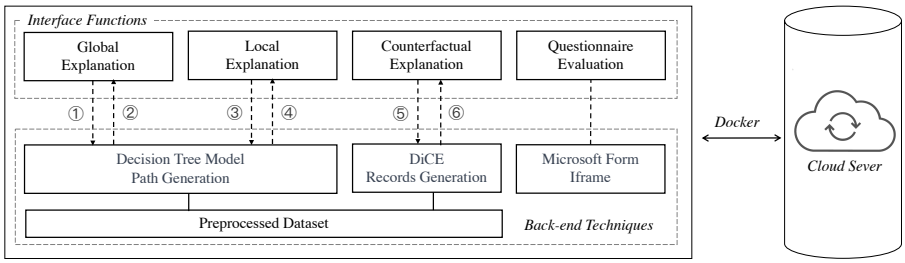


Fig. 2. System overview ((1) Maximum decision tree depth; (2) Decoded global path; (3) Maximum decision tree depth and selected record; (4) Decoded local path; (5) Original data sample and actionable counterfactual attributes; (6) Prediction and counterfactual data samples)

3.2 Dataset

Cardiovascular study[5] is a public data source with 4,237 records. Each record covers 15 attributes, containing information on people's demographic, behavioural, and medical status. We leverage these attributes to predict and explain whether individuals are at risk of developing coronary heart disease (CHD) within 10 years. Table 2 shows the dataset descriptions, where $\mathbb{Z}$ for Integer, $\mathbb{R}$ for real number and *Bool* for Boolean type.

The attributes *age, BMI, cigs/day, cholesterol, diaBP, sysBP, heart rate*, and *glucose* are categorized and encoded using OneHotEncoder. The remaining attributes are processed numerically using StandardScaler. Two of these processes are packaged in a ColumnTransformer and applied to the raw dataset for preprocessing.

[5] https://www.kaggle.com/datasets/christofel04/cardiovascular-study-dataset-predict-heart-disea

Table 2. Dataset Description

Attribute Name	Description	Value
Sex	Gender of the person	String: **M** for male, **F** for female
Age	Age of the person	$\mathbb{Z}$: An Integer ≥ 0
Is smoking	Whether a current smoker	Bool: 1 for true, 0 for false
Cigs Per Day	Average daily cigarette consumption	$\mathbb{Z}$: An Integer number ≥ 0
BP Meds	Whether on blood pressure medication	Bool: 1 for true, 0 for false
Prevalent Stroke	Whether had previously had a stroke	Bool: 1 for true, 0 for false
Prevalent Hyp	Whether was hypertensive	Bool: 1 for true, 0 for false
Diabetes	Whether had diabetes	Bool: 1 for true, 0 for false
Tot Chol	Total cholesterol level per deciliter	$\mathbb{Z}$: Normal ≤ 200 milligrams
Sys BP	Systolic blood pressure	$\mathbb{R}$: Hypertension ≥ 140 mmHg
Dia BP	Diastolic blood pressure	$\mathbb{R}$: Hypertension ≥ 100 mmHg
BMI	Body Mass Index	$\mathbb{R}$: Normal $18 \sim 25$
Heart Rate	Heart rate per minute	$\mathbb{Z}$: Normal $60 \sim 100$
Glucose	Glucose level	$\mathbb{Z}$: Normal $\leq 200 mg/dL$
10-year CHD	10-year risk of coronary heart disease	Bool: 1 for true, 0 for false

3.3 Visualisations of Global and Local Explanations

Decision trees and nodes of a tree produced directly by a decision tree algorithm contain lots of additional information, in addition to the attribute name and its split condition as seen in Figure 1. The additional information may include for example, GINI values or the number of records reaching a leaf nodes. Such information is difficult for non-expert users to understand and there is actually no need to present such information to non-expert users.

Our objective is to provide more straightforward explanations by visualising the decision tree. Figure 3 illustrates the process of converting text into an interactive graphical decision tree.

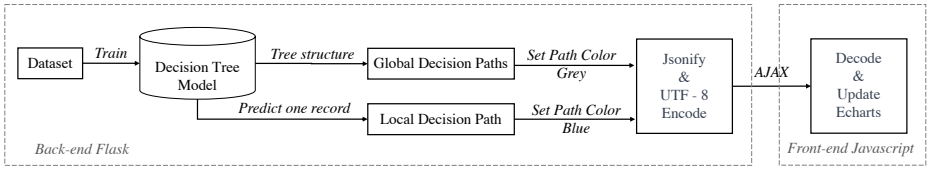


Fig. 3. The process of visualising decision trees

The decision tree model is based on implementation of CART in scikit-learn[7]. After training, the global decision path is formed by recursively exploring the `'tree_'` attribute in the model. The local decision path can be obtained by identifying a specific `'node_id'` before exploring the `'tree_'`. To differentiate global and local explanations, we set the global paths in a grey line style, while the local paths are shown in blue lines. After that, we convert the

path into a JSON format and encode it. The encoded string is then passed to the front-end by Flask[15] and Ajax. In the JavaScript of the HTML webpage, we utilise 'atob' and 'JSON.parse' functions to decode the data into a normal format. The decision tree is updated whenever we reset the Echarts[18] with the latest decoded data.

Visualization Optimization: Scalability and selective tree depths are implemented in the tree visualization to enhance user experience. The interface enables the generation of global explanations with a range of maximum tree depths between 2 and 6. Based on the tree we described in Section 3.2, Figure 4 shows global explanations with the tree depths of 2 and 3 respectively.

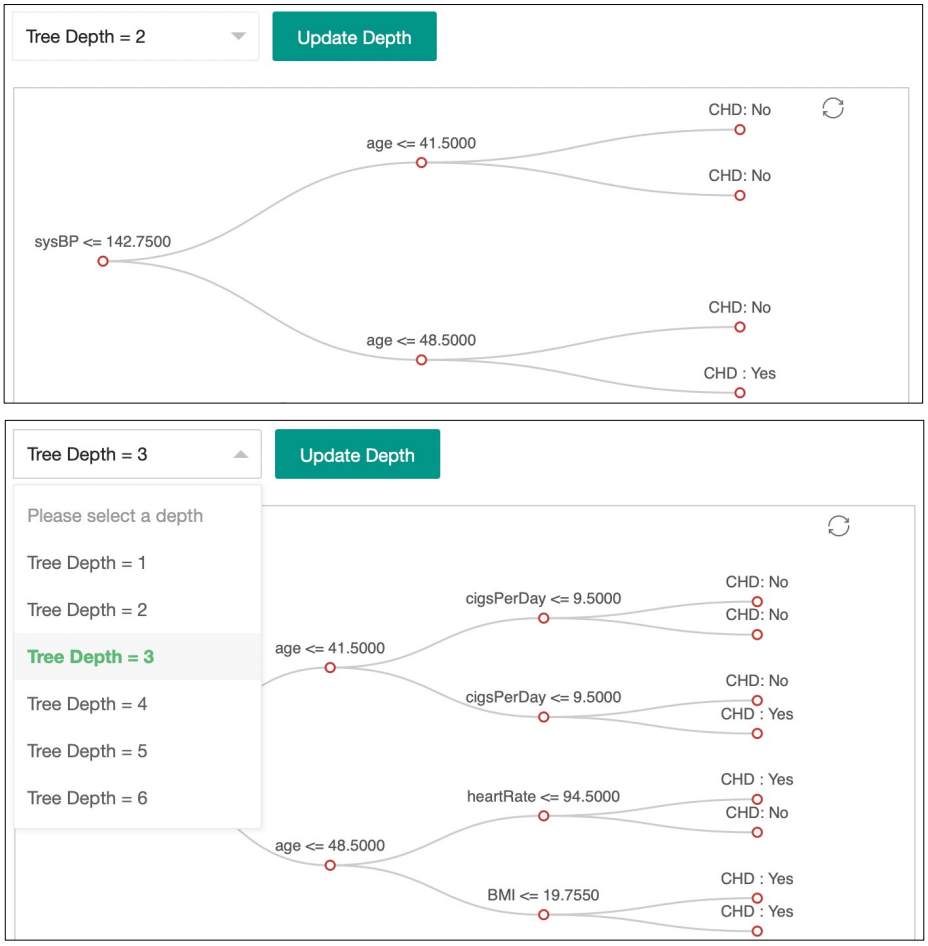


Fig. 4. The global explanation tree of different tree depths, where each node is expandable upon a click.

It's an interactive function that sends an Ajax request back to the model and returns the corresponding tree model when the user selects a depth and clicks the 'update depth' button.

The tree structure becomes more complex as the tree gets deeper. To enhance readability, we hide some nodes when a tree depth exceeds 4. Nodes can be expanded or hidden by clicking on a particular node. Figure 5 shows the tree zoom in and out function.

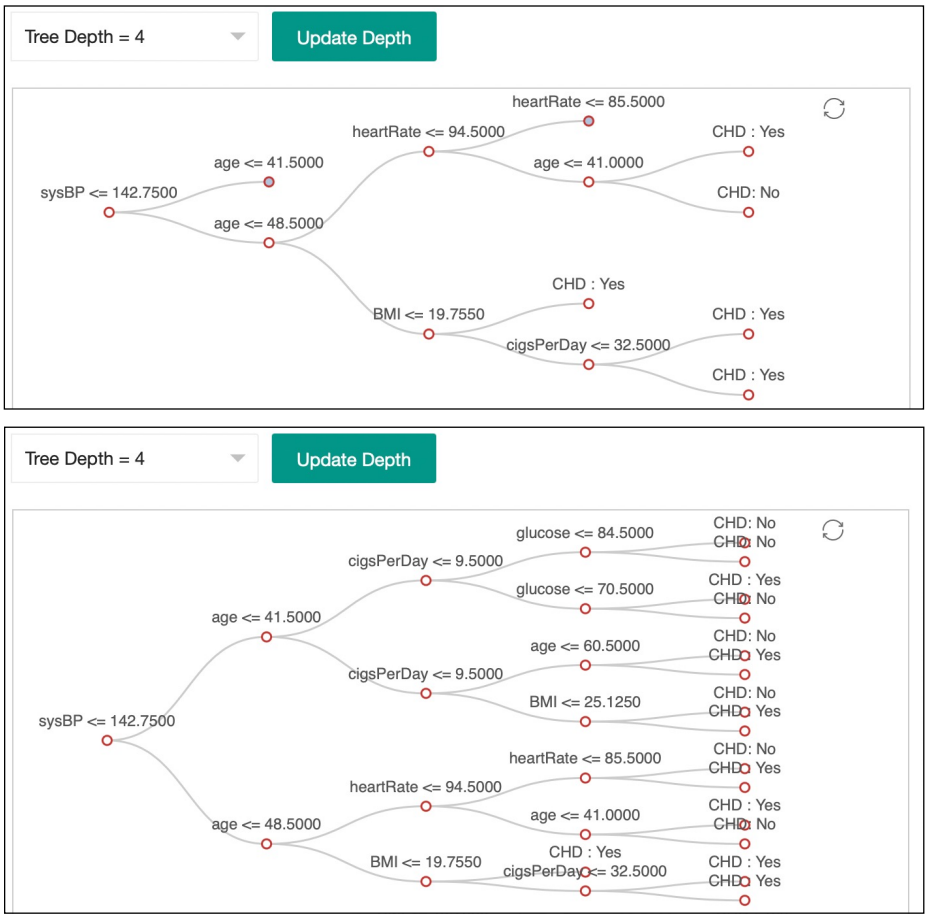


Fig. 5. Zoom in and out of explanation tree

3.4 Counterfactual Explanation Criteria

Changeable Attributes Selection: The dataset contains 15 attributes and attribute selection can reduce the dimensionality and enhance the computational efficiency of the algorithm. Moreover, some attributes are not actionable, for instance, an individual's age, or gender. Identifying and focusing on the most important attributes improves the actionability and interpretability of the counterfactual records. We use Shapley Additive Explanations(SHAP)[20] as a reference to select attributes.

SHAP is a method that quantifies attribute contributions for any machine learning model. SHAP calculates the Shapley value for each attribute given a specific record and measures its importance to the final outcome of this record. The intuition behind the Shapley value is to calculate the output difference with and without a specific attribute.

$$SHAP_{A(X_S)} = \sum_{S \subseteq F \setminus \{A\}} W * [f_{S \cup \{A\}}(X_{S \cup \{A\}}) - f_S(X_S)] \tag{7}$$

$$W = \frac{|S|!(|F| - |S| - 1)!}{|F|!} \tag{8}$$

In Equations 7 and 8, F is the set of all attributes in a dataset, A is an individual attribute, and $F \setminus \{A\}$ means the set of attributes without A. $f_{S \cup \{A\}}$ is a ML model with attribute A present, and f_S is a model without A, and X_S represents the values of attributes in S.

SHAP value for a single attribute of a particular record can be extended to calculate SHAP values for this attribute over all of the records in a dataset. The global importance of this attribute is then obtained by averaging these individual SHAP values for this attribute. The ranking of the attributes based on (global) SHAP values is shown in Figure 6.

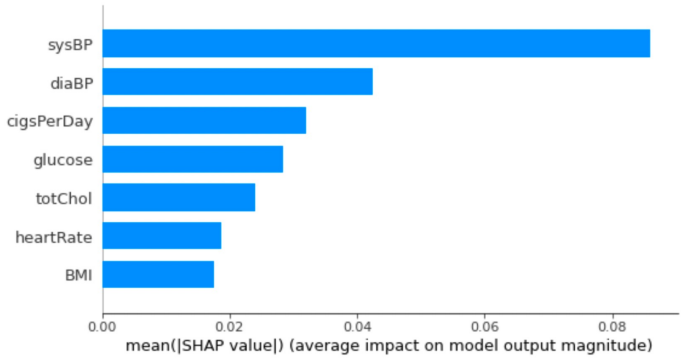


Fig. 6. The SHAP attribute importance plot

In practice, we first select important key attributes referring to the explanation of attributes contributions using SHAP. Then, we opt-out attributes that are not actionable, e.g., gender and age, not possible to change easily. Last, sysBP, diaBP, BMI, HR, cigs per day, glucose are 6 changeable attributes we provide when generating counterfactual records.

Attributes Range Constraints Range constraints are used for filtering potential infeasible counterfactual records due to real-world limits. Table 3 below lists range constraints of changeable attributes when generating counterfactual explanations.

Table 3. Attributes range constraints

Cigs/Day	Sys BP	Dia BP	Heart Rate	BMI	Glucose
[0, 400]	[0, 300]	[0, 250]	[0, 200]	[10, 50]	[0,250]

4 Interactive XAI System and its Evaluation

4.1 System Testing

The interface is available to access via weblink[6]. When users access the XAI interactive page for the first time, they will be redirected to the page containing a research introduction and participation ethical terms. Users must click the 'I consent' button to explore following main functions.

Data description, global and local explanation, counterfactual explanation, feedback & questionnaire are the four sections in the application. Most of the content on the data description page has been mentioned in Section 3.2.

In the following analysis, the terms **high risks**, **high likelihood**, and **more likely** are used interchangeably to refer to the 10-year CHD risk associated with a **Yes** result. Similarly, the terms **low risks**, **low likelihood**, and **less likely** refer to the 10-year CHD risk associated with a **No** result.

[6] https://med.bristol-xai.uk

Figure 7 shows the interface of global and local explanations, and illustrate the global and local explanations of a selected record (in this case, we set the tree depth as 3). Each of the 1011 records in the original dataset can be selected on the righ-panel to view its decision path displayed on the left-panel with blue colour, and the interface supports sorting by age and blood pressure in ascending or descending order.

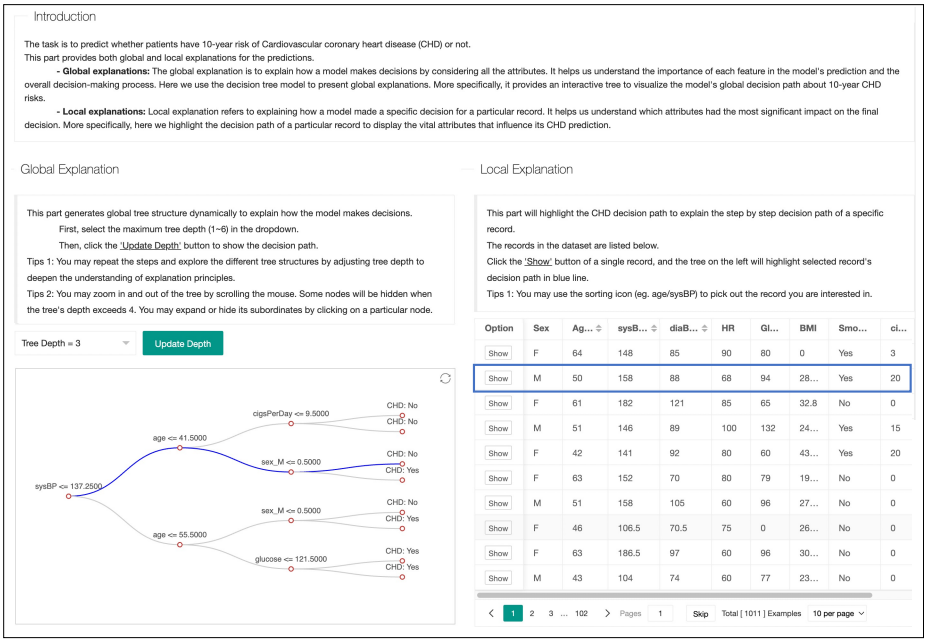


Fig. 7. Global and local explanation interface

Figure 8 presents the counterfactual explanation interface. A brief introduction to counterfactual explanation is provided at the top. Default values have been set to facilitate users in exploring the two panels below more conveniently.

Panel 1 displays how a 10-year CHD risk prediction can be generated when the button "Step 1" is clicked by a user. To aid a user in inputting these values, we provide some default values in these attribute boxes to start with. If a user wishes to alter any of these values, they can do so, and clicked the Step 1: Trigger ML Model for Prediction. The tabular display will appear beneath the button and show the prediction result (far-right in green colour).

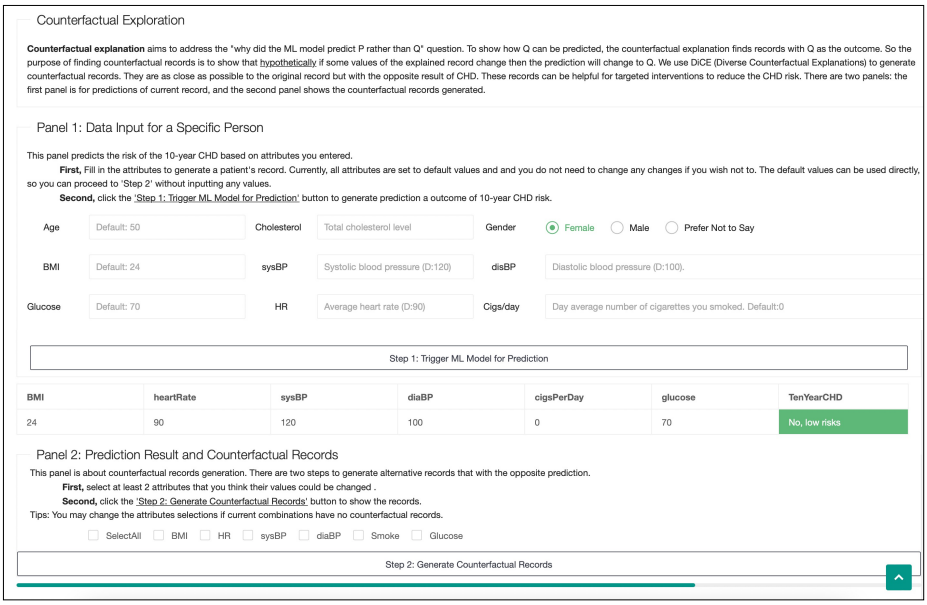


Fig. 8. Counterfactual explanation interface

For Panel 2, Figure 9 shows the counterfactual records found for the original record displayed in Panel 1. These counterfactual records are provided by using DiCE and considering the importance of attributes measured by SHAP given in Figure 6. In the counterfactual records, when the sign "-" is displayed under an attribute name, it means, the value of this attribute is the same as that in the original record. Counterfactual records only display attribute values which are different from original, and these values actually show how changes in some attribute values will contribute to generating a different prediction outcome. This is the essence of counterfactual explanation.

Fig. 9. Generate high risks CHD counterfactuals given a low-risk original record

Similarly, if the original record show "Yes, High-risk" and we want to provide counterfactuals to a user, "No, low-risk" records will be produced by DiCE to show how changes in some attribute values can alter the prediction outcome to "No" as shown in Figure 10. It shall be pointed out that in this case, only two attributes are used to find counterfactuals. Users are provided with a choice of how many attributes they wish to use to find counterfactuals. This is done by selecting attribute names in Panel 2, such as either "Select All", or just select some by clicking on individual attribute names.

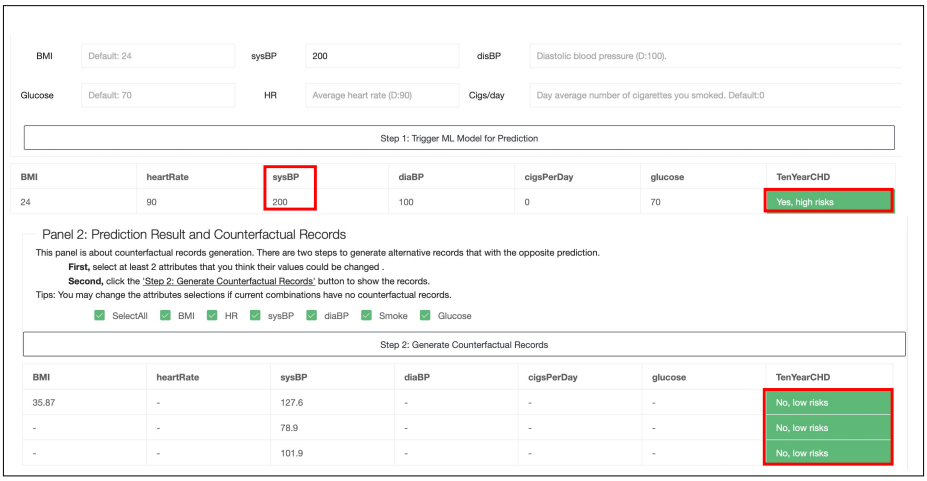


Fig. 10. Generate low risks CHD counterfactuals given a high-risk original record

Figure 11 shows an example of counterfactual records generated by using 4 attributes.

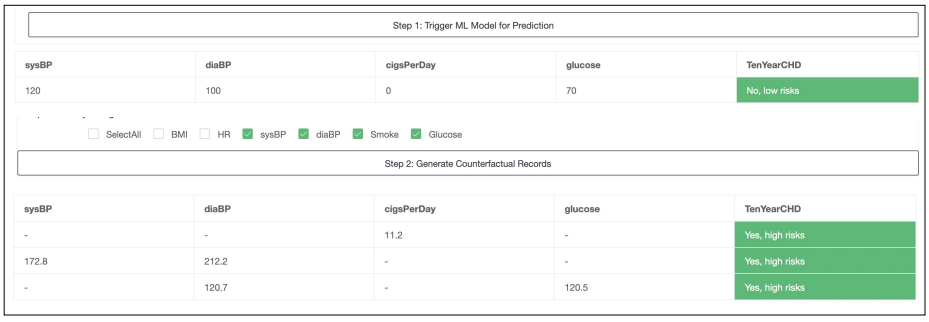


Fig. 11. Counterfactual explanations with 4 changeable attributes

When a counterfactual does not exist if a certain combination of attributes selected. When this happens, the system will prompt 'No Counterfactuals found for the given configuration, perhaps try with different attributes combinations. Recommend: try selectAll attributes for the first attempt'.

4.2 Questionnaire Evaluation

XAI aims to provide decision-makers with explanations of the computing system to help them understand that the entire process is reasonable. It is important to evaluate the effectiveness of the XAI explanation through a questionnaire for participants.

Table 4 shows our 14 questions covering three aspects: basic information, system satisfactory, and overall experience. The designed questions refer to the metrics for explainable AI proposed by [16], including goodness, satisfaction, and understanding.

Table 4. Questions for evaluation

Class	Alias	Question	Options
Basic Information	Q1	Your age	< 20; 20~30; 30~40; 40~50; >50
	Q2	Your gender	Female;Male;No binary;Prefer not to say
	Q3	Do you have any prior Artificial Intelligence experience?	Yes,a STEM related student/worker/researcher; No
	Q4	Do you have any prior Medical Domain experience?	Yes (a medical student/worker/researcher); Yes (been diagnosed with CHD/any related diseases); No (mainly focus on other domains)
	Q5	How do you describe your English proficiency?	Beginner, Intermediate, Advanced, Proficient, Native
System Satisfactory	Q6	Rank the explanations from easiest to least understandable in the list.	Data Description, Global Explaination, Local Explaination, Counterfactuals Explore
	Q7	How useful do you find the decision tree to your understanding the global & local explanations?	Five levels from very unhelpful to very helpful
	Q8	How useful do you find the counterfactual examples it generates to your understanding of counterfactual explanations?	Five levels from very unhelpful to very helpful
The Helpfulness of Explanations	Q9	How the decision tree works	Five levels from very unclear to very clear
	Q10	Why a certain prediction is given.	Five levels from very unclear to very clear
	Q11	How each attribute influences the result.	Five levels from very unclear to very clear
	Q12	How attributes combinations influence the result.	Five levels from very unclear to very clear
Overall Feedback	Q13	Do you encountered any challenges while using XAI	Optional
	Q14	Do you have any additional comments or feedback	Optional

Result Analysis: In total, 200 participants tried our XAI system after filtering out incomplete questionnaries. 34.5% under 20, 49.5% aged between 20 and 30, and 16% over 30 years old. As for English proficiency, 20% are beginners, 37% are intermediate, and 43% are advanced or native speakers. Additionally, there are 103 non-experts in AI and 120 individuals unfamiliar with healthcare.

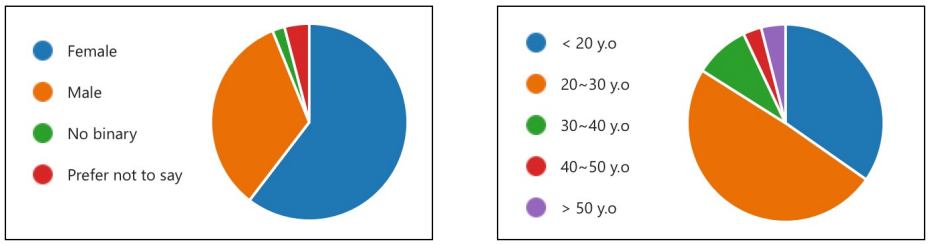


Fig. 12. The distributions of gender and age

Questions related to system satisfaction (Q7 and Q8) receive high ratings, Figure 13 shows that most of participants believe that the interfaces are helpful in understanding the decision tree classifier.

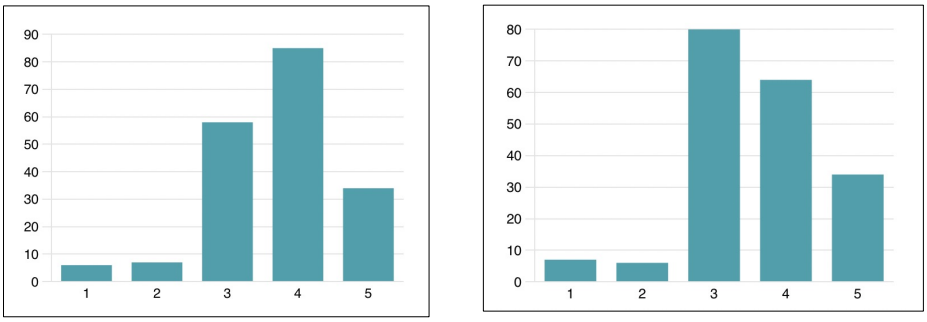


Fig. 13. Rates distributions of Q7(left) and Q8(right)

Table 5 reflects the mean, median, Standard Deviation(SD), Coefficient of Variation(CV) scores of question Q7 to Q12. The mean scores reflect the participants' overall ratings. The median values display central tendency and are robust to skewed distributions. SD and CV scores take into account the dispersion of rating data and can be used to assess the stability and consistency of the scores.

Table 5. Mean, median, SD, CV scores for Q7 to Q12

	Q7	Q8	Q9	Q10	Q11	Q12
Mean	3.705	3.586	3.490	3.469	3.367	3.376
Median	4.000	4.000	4.000	4.000	3.000	4.000
Standard Deviations	0.913	0.941	1.046	1.088	1.061	1.079
Coefficient Variation	24.64%	26.24%	29.97%	31.36%	31.51%	31.96%

Q7 gains highest average score of 3.705, and the lowest standard deviation of 0.913, suggesting that most participants agreed that decision trees are useful for understanding global and local explanations. Although Q8 receive a lower score, the average grade of 3.586 still suggests counterfactual explanations are effective. For the median values, most questions received a score of 4, while Q11 gained a median score of 3.

We further examined the coefficient of variation and found the variation in scores for Q10, Q11, and Q12 was above 31%. This indicates significant differences exist in how participants considered questions related to why a certain prediction is given (Q10), how each attribute influences the result (Q11), and how attribute combinations influence the result (Q12).

These findings show a positive attitude among participants towards our XAI interface but different satisfactory degrees exist for participants with different backgrounds.

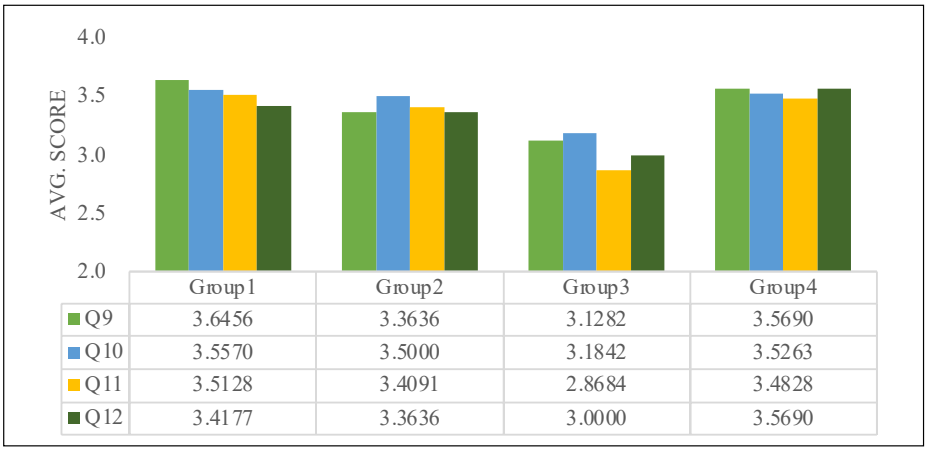
To further investigates the effect of participants' background, participants are divided into four groups based on whether they have prior knowledge of the AI field (Q3) and healthcare(Q4). Those with medical knowledge, such as medical students, medical workers, or those previously diagnosed with CHD-related diseases, are considered to have prior knowledge of healthcare, and the distribution for each group are shown in Table 6.

Table 6. Background group description and distribution

Group Names	**AI Field**	**Healthcare Field**	**Percentage(%)**
Group1	No	No	40.1%
Group2	No	Yes	11.2%
Group3	Yes	No	19.3%
Group4	Yes	Yes	29.4%

Figure 14 shows the scores given by each group for questions Q9 to Q12. Groups without AI-related knowledge (Group 1, 2) find the explanation useful and clear, with average scores around 3.5. One interesting finding is that participants with AI but no medical background (Group 3) give the lowest ratings, with average scores around 3.

Also, we read optional feedback (Q13,Q14) from Group 3 and found one possible reason for their critics is their familiarity with the ML model and expect more advanced implementations. They provide many helpful suggestions for further improvement including applying the technique to large-scale datasets, using figures like PDP plots to show attribute importance, and focusing on the robustness of the interface. The feedback from other groups is more general, including applying the current technology to other fields and providing support for different languages.



	Group1	Group2	Group3	Group4
Q9	3.6456	3.3636	3.1282	3.5690
Q10	3.5570	3.5000	3.1842	3.5263
Q11	3.5128	3.4091	2.8684	3.4828
Q12	3.4177	3.3636	3.0000	3.5690

Fig. 14. Comparison of Q9-Q12 Average Scores by Group

5 Conclusion

This paper presents an interactive web application that provides decision support to non-experts. The system offers intuitive global, local, and counterfactual explanations to visualise how a decision tree classifier works.

Compared to traditional static GUIs, the system provides dynamic explanations that enable users to (1) adjust the maximum depth of the decision tree, (2) personalize predictions based on textual inouts, and (3) generate counterfactuals based on different attribute combinations.

The system is applied in the healthcare field and evaluated through feedback from online participants. The results demonstrate that XAI methods can improve the model's credibility by helping users understand how and why it predicts a specific outcome. Moreover, users can deepen their understanding of the XAI system by experimenting with various inputs and observing changes in dynamic explanations Our work shares similarities with ExpliClas [2]. ExpliClas is a web service that generates global and local explanations after the user selects a dataset and a classifier. The main distinction is that we also offer counterfactual explanations with adjustable attributes. This feature offers a deeper insight into why the ML model predicted P instead of Q. Nonetheless, we are inspired to pursue the following improvements in future research: (1) support multiple datasets for user flexibility, (2) personalize user experience through their interactions with the interface, (3) track user browsing duration to assess the XAI system's attractiveness and (4) provide some additional explanations such as attribute-importance plot using DPP or LIME.

6 Acknowledge

This work is partially funded by the EPSRC CHAI project (EP/T026820/1)

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence. IEEE access **6**, 52138–52160 (2018)
2. Alonso, J.M., Bugarín, A.: Expliclas: automatic generation of explanations in natural language for weka classifiers. In: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–6. IEEE (2019)
3. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information fusion **58**, 82–115 (2020)
4. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 (2019)
5. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. IJCAI-17 workshop on explainable AI (XAI) **8**(1), 8–13 (2017)
6. Bove, C., Aigrain, J., Lesot, M.J., Tijus, C., Detyniecki, M.: Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In: Proceedings of 27th international conference on intelligent user interfaces. pp. 807–819 (2022)
7. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al.: Api design for machine learning software: experiences from the scikit-learn project. arXiv preprint arXiv:1309.0238 (2013)
8. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., Zhu, H.: Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders. In: Proceedings of the 2019 chi conference on human factors in computing systems. pp. 1–12 (2019)
9. Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable artificial intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **11**(1), e1391 (2021)
10. Craven, M., Shavlik, J.: Extracting tree-structured representations of trained networks. Advances in neural information processing systems **8** (1995)
11. Eifler, R., Brandao, M., Coles, A.J., Frank, J., Hoffmann, J.: Plan-property dependencies are useful: A user study. In: ICAPS 2021 Workshop on Explainable AI Planning (2021)
12. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Gregor, S., Benbasat, I.: Explanations from intelligent systems: Theoretical foundations and implications for practice. MIS quarterly pp. 497–530 (1999)
15. Grinberg, M.: Flask web development: developing web applications with python. " O'Reilly Media, Inc." (2018)
16. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable ai: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
17. Kulesza, A., Taskar, B., et al.: Determinantal point processes for machine learning. Foundations and Trends® in Machine Learning **5**(2–3), 123–286 (2012)

18. Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., Zu, M., Chen, W.: Echarts: a declarative framework for rapid construction of web-based visualization. Visual Informatics **2**(2), 136–146 (2018)
19. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
20. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017), http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
21. Martens, D., Provost, F.: Explaining data-driven document classifications. MIS quarterly **38**(1), 73–100 (2014)
22. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence **267**, 1–38 (2019)
23. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in ai. In: Proceedings of the conference on fairness, accountability, and transparency. pp. 279–288 (2019)
24. Molnar, C.: Interpretable Machine Learning. 2 edn. (2022), https://christophm.github.io/interpretable-ml-book
25. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P., Deepfool: A simple and accurate method to fool deep neural networks. In: Proceedings of the CVPR. pp. 2574–2582
26. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
27. Myles, A.J., Feudale, R.N., Liu, Y., Woody, N.A., Brown, S.D.: An introduction to decision tree modeling. Journal of Chemometrics: A Journal of the Chemometrics Society **18**(6), 275–285 (2004)
28. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
29. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
30. Stevenson, A.: Oxford dictionary of English. Oxford University Press, USA (2010)
31. Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I.: An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine **3**(1), 17 (2020)
32. Swartout, W.R.: Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: Bg buchanan and eh shortliffe,(addison-wesley, reading, ma, 1984); 702 pages, 40.50 (1985)
33. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL Tech. **31**, 841 (2017)