FeatureCloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Deliverable D4.8
"Explanation strategies, i.e. post-hoc vs. ante-hoc approaches"

Work Package WP4
"Supervised Federated Machine Learning"

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

© FeatureCloud Consortium, 2023
This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both. Reproduction is authorised provided the source is acknowledged.

Document information

Grant Agreement Number: 826078				**Acronym: FeatureCloud**
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Christos MARAMIS, Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Jan Baumbach, University of Hamburg (UHAM)			
Deliverable	D4.8 - Explanation strategies, i.e. post-hoc vs. ante-hoc approaches			
Work Package	WP4 - Supervised Federated Machine Learning			
Date of Delivery	**Contractual**	30/06/2023 (M54)	**Actual**	29/06/2023 (M54)
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	03 MUG			
Responsible Author(s)	Andreas Holzinger (AH), Alessa Angerschmid (AA), Anna Saranti (AS), David Schneeberger (DS)			
Keywords	Artificial Intelligence (AI), explainable AI, XAI, Post-Hoc, Ante-Hoc, Explanations			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	15/04/2023	First Draft	AH (MUG), AA (MUG), DS (MUG)
V0.2	01/05/2023	Comments	AH (MUG), DS (MUG), AS (MUG)
V0.3	15/05/2023	Usability Evaluation	AA (MUG)
V0.4	26/05/2023	Comments / Experiences	DS (MUG), AH (MUG)
V0.5	01/062023	Revised Draft	AH (MUG), AA (MUG), AS (MUG), DS (MUG)
V0.6	10/06/2023	Usability Results	AA (MUG)
V0.7	16/06/2023	Reviewable Version	AH (MUG), AA (MUG), AS (MUG), DS (MUG)
V0.8	28/06/2023	Reviewed Version	Internal Review and Approval by Jan Baumbach (UHAM)
V1.0	29/06/2023	Submission-Version	Nina Donner (concentris)

Actual effort in person-months (PMs)

Contributor (name and institution)	Invested resources (deliverable)	Overview of contributions
Holzinger, Andreas (MUG)	0.50 PM	First draft, comments, draft, adjustments, final version
Saranti, Anna (MUG)	0.25 PM	First draft, comments, draft, adjustments, final version
Angerschmid, Alessa (MUG)	0.25 PM	First draft, comments, draft, adjustments, final version
Schneeberger, David (MUG)	0.25 PM	First draft, comments, draft, adjustments, final version
Jan Baumbach (UHAM)	0.05 PM*	Review, feedback, approval
Nina Donner (concentris)	0.05 PM	Formatting and submission

*This person dedicated a certain amount of time to FeatureCloud, but received no salary from the FeatureCloud budget (e.g. Professor, Supervisor, Intern, Master/Bachelor student, etc.).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Table of acronyms and definitions

AI	Artificial Intelligence
AIA	Artificial Intelligence Act
Art.	Article
BRL	Bayesian Rule List
CLARUS	InteraCtive ExpLainable PlAtform for GRaph NeUral NetworkS
concentris	concentris research management gmbh
D	Deliverable
DIN	Deutsches Institut für Normung
DKE	Deutsche Kommission für Elektrotechnik Elektronik Informationstechnik
DoA	Description of Action
DT	Decision Tree
GAM	Generalized Additive Model
GDPR	General Data Protection Regulation
GND	Gnome Design SRL
GNN	Graph Neural Network
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
ISO	International Organization for Standardization
MS	Milestone
MUG	Medizinische Universität Graz (Medical University Graz)
NN	Neural Network
para.	paragraph
Patients	In this deliverable, we use the term “patients” for all research “subjects”. In FeatureCloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus, to increase readability, we simply refer to them as “patients”.
PPI	Protein-Protein-Interaction
RQ	Research Question
SBA	SBA Research Gemeinnützige GmbH
SCS	System Causability Scale
SDO	Standards Development Organization
SHAP	SHapley Additive exPlanation
SUS	System Usability Scale
WP	Work package
XAI	Explainable Artificial Intelligence

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

2 Objectives of the deliverable based on the Description of Action (DoA)

The objective of this deliverable is based on the DoA, which incorporates mostly aspects from Objective 6: “to find the best suitable explanation strategies, i.e., post-hoc and ante-hoc approaches and testing the user interpretation on the demonstrator in order to redesign the ‘explanation interface’” (Tasks 6 and 7). In this technical report **D4.8 “Explanation strategies, i.e. post-hoc vs. ante-hoc approaches”**, MUG presents the work carried out together with his partners within Task 6 (the evaluation of CLARUS: An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks), and particularly task 7 “Experiment and evaluate different explanation strategies” including the results of the evaluation and usability tests of CLARUS. This also fulfils **MS30 “Explanations Strategies for the human-in-the-loop”**.

3 Executive Summary

This deliverable D 4.8 “Explanation strategies, i.e. post-hoc vs. ante-hoc approaches” is designed to provide an in-depth exploration of explanation strategies, which is of rising importance in medicine and beyond. The essence is in Explainable Artificial Intelligence (XAI) which aims to make “black-box” machine learning results transparent, re-traceable, re-enactable and eventually understandable, and the various methods and strategies that can be employed to achieve it.

This report is composed of three distinct parts (A, B, C), each of which provides valuable insights and perspectives on this important and rapidly evolving field and which shall be helpful for the international research community.

In Part A, we delve into the world of standardisation and regulatory activities and their impact on Explainable AI. We explore the various frameworks and guidelines that have been developed to promote transparency and accountability in AI systems and examine their relevance and applicability to the field of Explainable AI. We also consider the various challenges and opportunities that arise in the context of regulatory oversight and discuss the potential benefits that can accrue from a well-regulated AI landscape.

In Part B, on the other hand, we focus on the different methods and strategies that can be employed to achieve Explainable AI. Specifically, we examine the two broad categories of Post-hoc and Ante-hoc methods and compare and contrast their respective strengths and weaknesses. We also explore the various explanation strategies that have been developed to improve the interpretability and transparency of AI systems and highlight the important role that measurements and metrics play in the evaluation and comparison of these strategies. An overview of the various open-source implementations of these Post-hoc and Ante-hoc methods that are available to researchers and practitioners in the field of Explainable AI is provided. We discuss the advantages and limitations of these implementations and highlight some of the key features and functionalities that operators should be aware of.

In Part C, we investigate the usefulness of the interactive XAI prototype CLARUS, presented in D4.6. An online user study was conducted to examine the usability and interpretability of the system. CLARUS allows experts to observe how changes based on their questions affect the AI decision and the corresponding XAI explanation. This tool was developed to counteract the lack of trust in artificial intelligence (AI) models in medicine and enable their use in clinical decision support systems (CDSS). The interactive XAI platform facilitates domain experts to ask manual counterfactual ("what-if") questions. As a result, these explanations should lead to some degree of causal understanding by a clinician in the context of a specific application.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

4 Introduction (Challenge)

In our work we faced three different, but highly interlinked challenges (A, B, C) related to "Explanation strategies, i.e., post-hoc vs. ante-hoc approaches" all related to the field of explainable Artificial Intelligence (XAI):

The first challenge (A) is directly related with explanation strategies, particularly with the impact of standardisation and regulatory activities. This challenge addresses the "side effects" of the expansion of the research domain XAI, which has become a significant counterbalancing force to the widespread adoption of complex black box models. This includes the currently so popular and highly successful large language models, based on the success of transformers and currently implemented in e.g., LLaMA, Alpaca, LaMDA/Bard, Chinchilla, ChatGPT, GPT-4, etc.

This incredible success, the wide-spread distribution and worldwide acceptance of AI have led to a proliferation of terminology and an array of diverse definitions, making it increasingly challenging to maintain coherence (Cabitza et al., 2023).

XAI refers to the development of AI systems that can provide clear, understandable, and interpretable explanations for their advice and decisions. The need for XAI has arisen because recent sub-symbolic approaches (such as ensemble methods or deep neural networks) have made machine learning approaches so complex, so high-dimensional, and so nonlinear that it has become very difficult to re-trace, re-enact and/or interpret results and represent them in a way that is understandable to humans, hence such approaches are referred to as black-box approaches (Goldstein et al., 2015), (Vidovic et al., 2015), (Castelvecchi, 2016), (Lakkaraju et al., 2017), (Guidotti et al., 2019), (Lakkaraju et al., 2019). And it is this challenge that forms the basis for XAI, which is now widely recognized as an essential aspect for the practical implementation of AI models (Arrieta et al., 2020). However, the very definition of explanation, and its mentioned desirable properties, is often not straightforward from a scientific point of view.

This scenario threatens to create a "tower of Babel" effect (the biblical story from Genesis which describes how humanity, speaking a single language, sought to build a tower to reach the heavens, leading God to create multiple languages, causing confusion and scattering people across the earth). This "tower of Babel" nowadays also creates a multitude of languages concerning XAI and impedes the establishment of a common (scientific) ground. In response to this enormous challenge, we respond within this project so that we first examine the quest for standardised definitions in the realms of standardisation and legislation where - in contrast to the scientific domain - a community-based agreement about central terms must be established as different definitions cannot co-exist. Subsequently, we propose a methodology for identifying a unified lexicon from a scientific standpoint (Schneeberger et al., 2023).

In the second challenge (B), i.e., Post-Hoc vs. Ante-Hoc Explanation Strategies - Experimental Evaluation, we analyse an interrelated problem. Because of the growth of XAI, which has become a large and complicated field with many different approaches, this has led to a widening gap between theory and practice, making it more cumbersome for non-expert data scientists to decide which method to use. Consequently, we present in our report a set of best practices, guidelines and templates for developing AI systems that are transparent, understandable, and explainable. Our guidelines are intended to help AI designers and developers create systems where users can understand, interpret, and comprehend how an outcome was arrived at (Retzlaff et al., 2023).

Finally, in the third challenge (C) we carried out an evaluation of CLARUS (short for An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks), which has been technically described in more detail in D4.6. For this to be useful, it is essential that the domain

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

experts are not only able to work with it (usability), but also to understand the explanations (causability).

Following the ISO 9241-11:2018EN, usability, as a measure for the quality of use (Bevan, 1995), is defined as the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Consequently, in our context of CLARUS, the challenge is to provide the end users (domain experts) with an interface that allows them to achieve their goals of investigating counterfactuals regarding the protein-protein interaction networks while being effective, efficient and satisfactory.

We introduced “causability” (Holzinger et al., 2019a) in reference to the well-known term “usability”. Whilst i) usability is about the usage of a system, and ii) explainability in the sense of XAI is about implementing transparency and traceability (Holzinger et al., 2022), so it is about a technical solution of highlighting what parts of the input contributed to a specific result, iii) Causability is about the measurement of the quality of explanations, i.e., the measurable extent to which an explanation of a statement to a user achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use (Holzinger et al., 2020).

Explainability technically highlights decision relevant parts of machine representations and machine models i.e., parts which contributed to model accuracy in training, or to a specific prediction. A good example is pixel-wise relevance propagation (Bach et al., 2015) or layer-wise relevance propagation (Montavon et al., 2019).

However, this explainability does NOT refer to a human model, while causability explicitly refers to the human model (Holzinger, 2020), (Holzinger et al., 2021), (Plass et al., 2023). Successful mapping between explainability and causability requires new human-AI interfaces that enable contextual understanding and allow the domain expert to ask questions and be counterfactual ("what-if"-questions) and CLARUS is a good example for this approach (Beinecke et al., 2022). Critics of this approach keep asking what the human-in-the-loop should do; the human-in-the-loop (Holzinger, 2016), (Holzinger et al., 2019b) can (sometimes – of course not always) bring human experience and conceptual knowledge to AI processes - something that the best AI algorithms on the planet still are lacking (Zador et al., 2023).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

5 Methodology

5.1 Part A: Explanation Strategies - impact of Standardization and regulatory activities

The goal of this part of the D4.8 was to analyse the interaction between the ongoing standardisation and legal efforts concerning XAI and the scientific perspective on the other side. In addition, the goal was to show possible frictions between those two perspectives and to propose a solution to the described “Tower of Babel” problem. Therefore, we mapped the current efforts concerning the standardisation of transparency and XAI by conducting a review of existing standards and standards in the drafting stage. Secondly, a legal analysis incorporating recent scholarship and jurisprudence was used to map the role of the General Data Protection Regulation and the Artificial Intelligence Act in regulating XAI. The proposal of a communal initiative to define XAI terms draws on existing scoping XAI reviews and comparable communal efforts in finding common definitions.

5.2 Part B: Post-Hoc vs. Ante-Hoc Explanation Strategies - Experimental Evaluation

The goal of this part of the deliverable was to evaluate the implementation of post-hoc vs ante-hoc explanations. In order to verify the factors and the correct assessment for each of the XAI approaches, expert interviews were conducted. The study uses a within-subject design, i.e. each participant is exposed to each condition (Baxter and Jack, 2008), whereby condition is referring to the different explanations. Thus, each participant is presented with each explanation and is able to compare them.

The Iris dataset was chosen as an example dataset to create the explanations for the experiments because firstly a broader target audience can be reached since the data set is mostly self-explaining and not too complicated and secondly most software developers are familiar with text-based information. Thirdly, an easily understandable dataset allows test persons to focus on the task at hand and not the dataset itself.

5.2.1 XAI background

The terms "explainability" and "interpretability" are often used interchangeably, but there are subtle differences between the two terms. Interpretability refers to the extent to which a human can understand the reason for a decision made by a machine learning model in the first place. It refers to the transparency of the internal mechanisms of a model. For example, an interpretable model allows a developer to understand the process the model uses to get from input variables to output prediction. This might include understanding which features are most important or how changes in those features might change the outcome. Linear regression models, for example, are often considered highly interpretable because the weight of each feature is explicitly modelled as a coefficient.

Decision trees also theoretically offer high interpretability, since the path from root to leaf provides a clear explanation of the decision process (however, large decision trees can also be very difficult to interpret, but the possibility just exists).

Explainability, on the other hand, is about providing understandable explanations for the decisions made by the machine learning model. Explainability is often associated with post-hoc interpretability, where methods are applied after the model has been trained to generate explanations.

There is a major difference in ante-hoc and post-hoc, namely when and how interpretability is built into the model. In ante-hoc interpretability (hence also called intrinsic interpretability), interpretability

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

is built into the model at the model development stage. The main goal here is to construct a model that is intrinsically interpretable. As mentioned above, decision trees, linear regression, and logistic regression are examples of such ante-hoc methods.

Post-hoc methods, on the other hand, are applied after the model has been trained. They are most often used with complex models (such as Deep Learning models) that are so complex and so high-dimensional that they are not inherently interpretable. These methods aim to provide an explanation for the behaviour of the model after it has been trained, without affecting the model architecture itself. In Table 1 we present three ante-hoc and three post-hoc methods.

Table 1: Comparison of post-hoc and ante-hoc methods.

	Performance (Fast Computation)	Fidelity/Correctness	Completeness	Time required
Feature Importance (SHAP) (Lundberg & Lee, 2017)	High	Medium	High	Low
Rule-Based XAI (Anchors) (Ribeiro et al., 2018)	Low (grows with features)	High	Medium	High
Counterfactuals (Dandl et al., 2020)	High	Medium	Medium	High
Decision Trees (DT) (Safavian & Landgrebe, 1991)	High	Medium	Medium	Low
Bayesian Rule Lists (BRL) (Letham et al., 2015)	Low	High	High	High
GAM (Generalised Additive Models) (Hastie, 1992)	Medium	High	High	Medium

The post-hoc methods used as comparison objects were:

1. SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) is one example of a feature importance technique. It is derived from game theory to measure feature importance and explain individual decisions i.e. it computes the contribution of each feature (Molnar, 2022).

2. Rule-based XAI (Anchors) (Ribeiro et al., 2018) are using a perturbation-based approach to explain individual decisions by finding a decision rule that “anchors” the prediction. I.e. a rule anchors a prediction if changes in other feature values do not affect the prediction. These anchors are presented as IF-THEN statements (Molnar, 2022).

3. Counterfactuals (Dandl et al., 2020) are a contrastive technique to explain individual predictions by describing the smallest change to the feature values that changes the prediction.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

The ante-hoc methods used as comparison objects were:

1. Decision Trees (DT) (Safavian and Landgrebe, 1991) split the data multiple times according to certain cutoff values in the features, creating different subsets of the dataset with each instance belonging to one subset. These models can be interpreted by following the tree structure, starting from the root node, along the next nodes and edges till the leaf node with the predicted outcome is reached (Molnar, 2022).

2. Bayesian Rule Lists (BRL) (Letham et al., 2015) are a type of a generative model which consists of decision lists with a series of IF-THEN statements which are interpretable by human experts as a high-dimensional multivariate feature space can be transferred into a low-dimensional and thus human-interpretable decision space.

3. Generalised Additive Models (GAMs) (Hastie, 1992) extend linear and logistic regression with the capability of modelling non-linear relationships. These models are useful in terms of understandability as long as low-dimensional terms are considered.

5.2.2 Interview Process

The interview consisted of two parts. Participants were presented with the generated explanations for a specific test case, as well as the assessment of the chosen factors for each approach. First, the experts are asked to rate the quality of the generated explanations on a 6-point Likert scale. The used Likert scale ranges from 1 (strongly disagree) to 6 (strongly agree), this allows increased measurement precision while forbidding a neutral option (Nemoto and Beglar, 2014). These interviews were used to verify the assessment and give further insight into the estimates of a data scientist.

The following quality metrics were used to compare the different XAI approaches:

1. completeness
    a. Did the explanations cover the entire model behaviour?
    b. Were there any important aspects of the model's behaviour or predictions that were not captured by the XAI method?

2. understandability
    a. Were there any limitations or difficulties in interpreting the explanations provided by the XAI method?
    b. How much effort and time did it take you to understand the explanation?
    c. What hindered you and what helped you?

3. appropriateness for the target audience
    a. Who is the intended audience for the explanations generated by the XAI method?
    b. Are there any specific user groups that would benefit more from the explanations than others?

Afterwards, the experts are asked to evaluate the approaches by answering the following more open questions for the general evaluation:

1. Do you have any previous use cases where you have employed XAI methods and what was the outcome? Did you have any specific XAI methods that you have found particularly effective in the past?

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

2. How important is it for you to have a complete understanding of how your machine learning model makes predictions? Do you remember cases where interpretability of results was especially important?

3. Do you need to instantly generate and regenerate explanations for your model? What is the maximum waiting time you could tolerate for generating an explanation?

4. Can you explain the reasoning behind the predictions generated by your machine learning model? What are the most important features that contribute to the predictions?

5. What would you look for in an explanation of your model? Which aspects do you want to understand most?

6. How large are the datasets you usually handle, and how complex are your models? To what number of features and data points should the method scale?

7. Have you encountered any issues with errors, outliers, or missing values in your data? How important is it for an XAI method to be robust in these scenarios?

8. How would you prefer to integrate an XAI method into your existing workflow? Are there any specific features or functionalities that you would require in an XAI tool to make it usable for you?

9. At whom would the explanations most likely be aimed? Will you use them to make the developers, management, and/or end users understand the model?

5.3 Part C: Usability Evaluation of CLARUS

In (Beinecke et al., 2022), we describe CLARUS, which is a system in which a human expert can interact as a human-in-the-loop and explore how changes based on his questions affect the AI decision and the corresponding XAI explanation. This interactive XAI platform allows the domain expert to manually ask counterfactual ("what-if") questions. With CLARUS, the expert can observe how the changes based on his questions affect the AI decision and the corresponding XAI declaration. The interactive XAI platform prototype allows not only the evaluation of specific human counterfactual questions based on user-defined alterations of sample graphs and a re-prediction of classes but also a retraining of the entire graph neural network after changing the underlying graph structures.

In order to connect the transparency of XAI models with the need for causability, we need interactive XAI platforms that guide domain experts through the explanations given by an XAI model. Such platforms will allow the expert to gain insight into what influenced the AI model during the decision-making process. Additionally, it can help the expert identify confounding factors that might inhibit model performance or correlated factors that are biologically unimportant. Most importantly, it is imperative to allow the expert to interactively ask counterfactual questions ("if-not", "why-not", and "what-if" questions) and change their data based on those. Ultimately, this interaction will help the expert see how changes affect not only the AI model but also the XAI model, hence, increasing their causal understanding. CLARUS will pave the way for informed medical decision-making and the application of AI models as clinical decision support systems (CDSS). The aim of CLARUS is to promote human understanding of GNN predictions and to allow the domain expert to validate and improve the GNN decision-making process. Therefore, we have developed a platform that visualises the input graphs used to train and test the GNN, including node and edge attributes, as well as, node and edge relevances computed by XAI models, such as GNNExplainer (Ying et al., 2019). Figure 1

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

shows an overview of the CLARUS interface. The Homepage illustrates the purpose and how to use the platform. In the 'Select Data' tab the user can choose between three pre-selected datasets, namely the Kidney Renal Clear Cell Carcinoma (KIRC) dataset (a real-world dataset) and a smaller subset of the KIRC dataset, as well as, a synthetic dataset. The synthetic dataset is built out of 1000 Barabasi networks with two synthetic classes. This dataset is smaller with only 30 nodes and 29 edges per graph making it easier for the user to get familiar with CLARUS than with the KIRC dataset. The KIRC dataset is a larger real-world dataset, taken from The Cancer Genome Atlas (TCGA) database (https://cancergenome.nih.gov/).

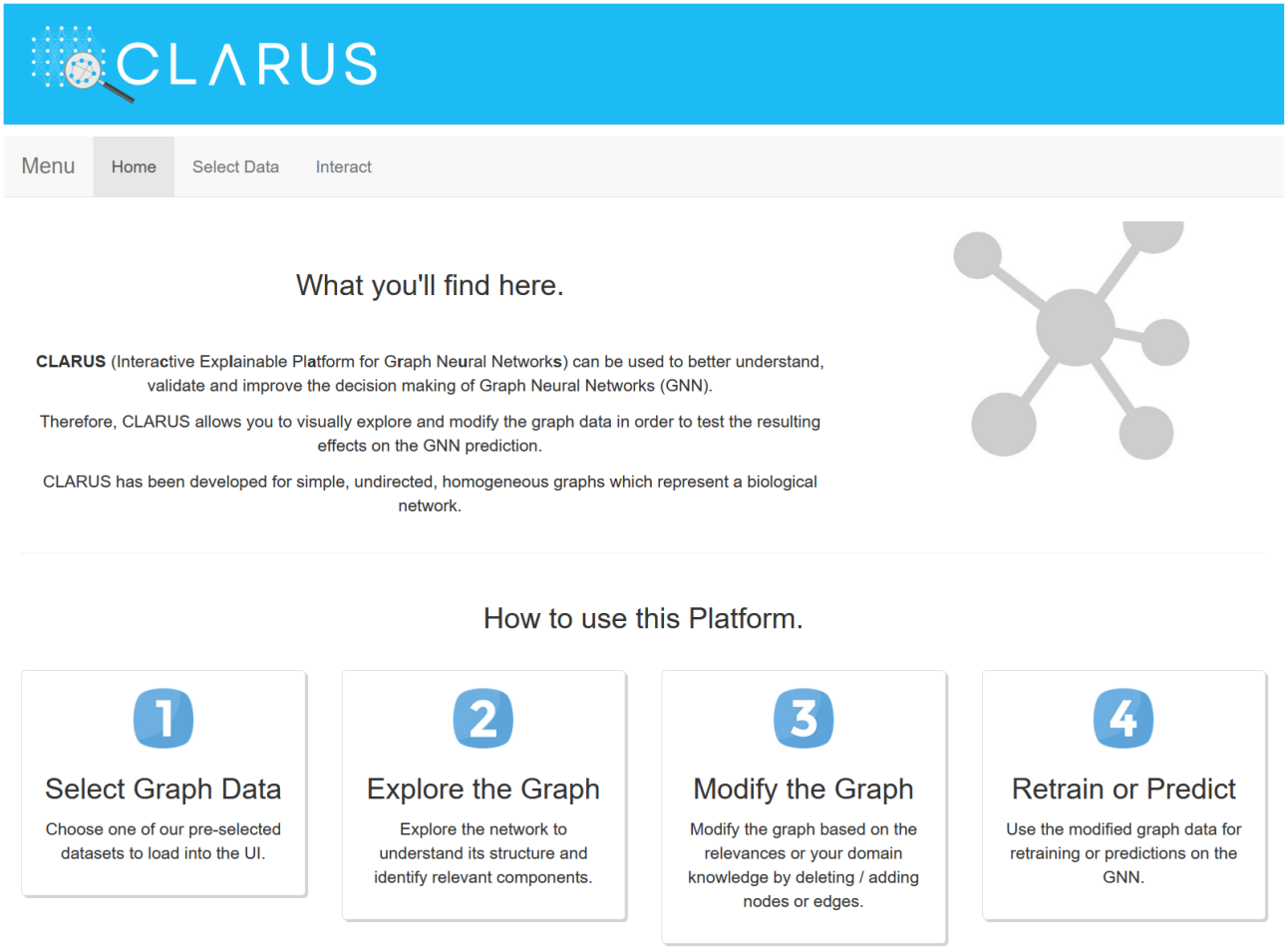

Figure 1: CLARUS Start page.

Figure 2 shows the interaction window with its seven main interaction components.

The first component is a drop-down menu that allows the user to select a graph from the dataset and gives the user some first information on the graphs, namely if the graph was used in the training or in the testing of the GNN, what class the graph belongs to, what class the GNN classified the graph as, and the confidence of the GNN's prediction (e.g. probability for a certain class). This enables the user to decide if they want to work on wrongly classified graphs or on training graphs to see if they can improve the GNN's performance.

The second component shows the confusion matrix, sensitivity and specificity of the GNN on the test-set data. After the graph selection, the third component allows the user to specify the graph visualisation. Here the user can select how to sort and colour the nodes, by degree, by relevances from XAI models or alphabetically according to their labels. Additionally, the XAI model used for edge

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

colouring can be selected.

Further, it can be specified how many nodes should be displayed in the graph visualisation. The graph visualisation can be found in the fourth interaction component and it displays the graph based on the user selections alongside a legend for the node colours. The fifth component consists of information on the edges and nodes displayed in the graph visualisation. This allows the user to get a more detailed view of the graph components. To ask counterfactual questions, the sixth component allows the user to manipulate the graph by adding or deleting nodes or edges. The last component is a log print, that shows the user all of his performed actions.

The system was evaluated by testing a disease module using GNN-SubNet (Pfeifer et al., 2022) to CLARUS in order to conduct an in-depth investigation of the performance and the distribution of the associated relevance scores. The resulting subnetwork or subgraph consists of four proteins, namely MGAT5B, MGAT5, MGAT4B, and MGAT3, connected by five edges. This was uploaded to CLARUS. Overall, the module had a high sensitivity of 0.95 and a specificity of 0.31 on an independent test data set (127 patients).

A patient who was classified with a high predictive confidence of 1.28 was selected in favour of the kidney cancer-specific class. The GNNexplainer assigned the highest relevance score to the node associated with the proteins MGAT3 and MGAT4B (GNNexplainer relevance score of 0.86). The edge between MGATT3 and MGAT5 had the highest saliency value of 0.98. The first manual counterfactual aimed at evaluating the importance of the edge with the highest relevance score MGAT3-MGAT5. The deletion of this edge led to a confidence drop of the predicted class to 0.98. Moreover, a notably higher importance was now assigned to the edge MGAT5-MGAT4B with a saliency value of 1. Interestingly, node importance was the same for MGATT3 and MGAT4B with a value of 0.86.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

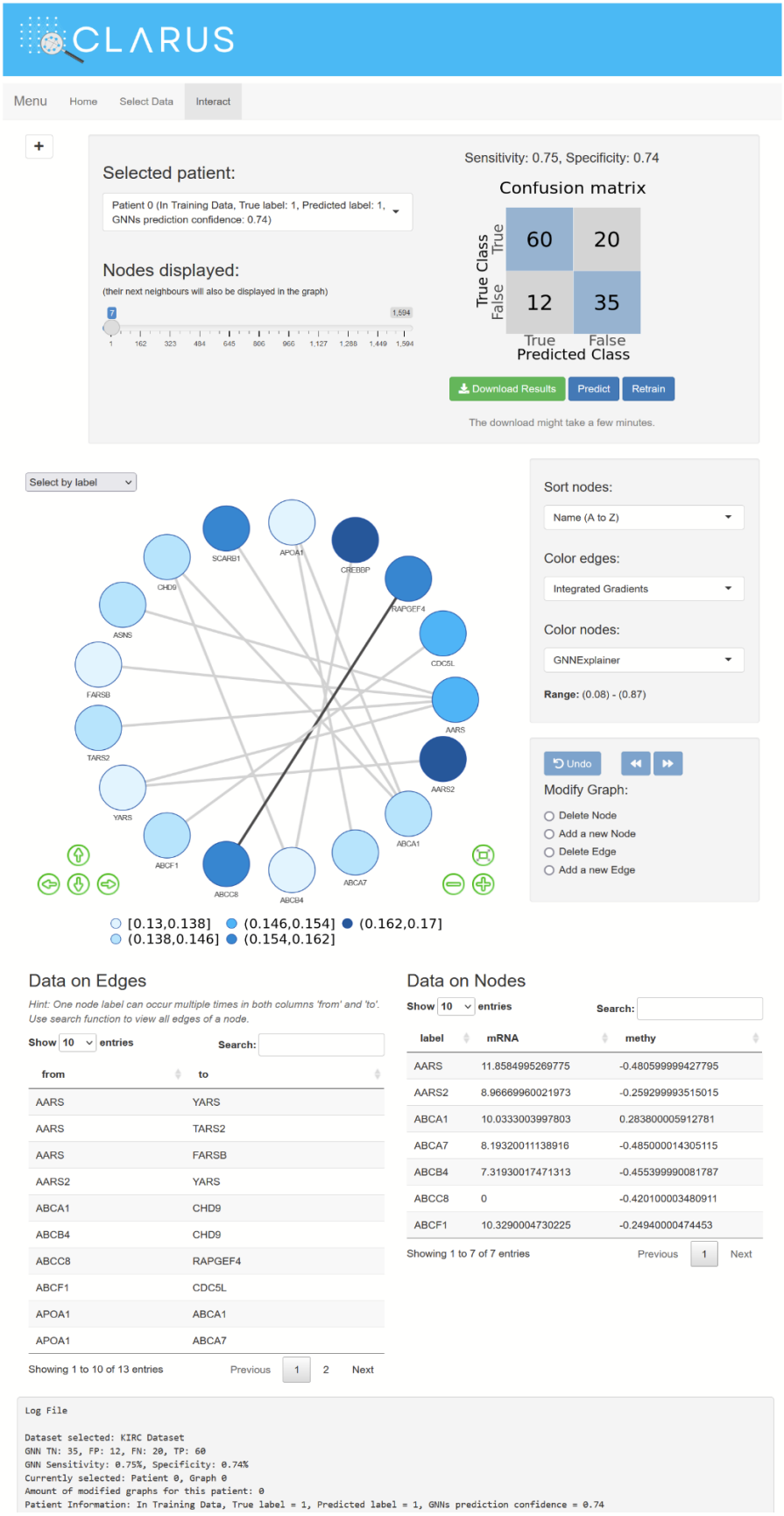

Figure 2: CLARUS interface overview.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The obtained results may suggest that MGAT3 and MGAT4B are redundant. However, given the remaining topology of the graph, it may be assumed that MGAT4B is more important due to a higher degree than the other nodes. Following this assumption, our next manual counterfactual was the deletion of the MGAT3 node. As a result, the prediction confidence increased again to 1.24 and the node importance of MGAT4B was still at 0.85. The importance of the MGAT5-MGAT4B edge was the same with a saliency value of 1. The edge between MGAT5 and MGAT4B was assumed to be the most relevant part of the analysed module.

According to this assumption and to validate our hypothesis we further deleted the edge between MGAT4B and MGAT5, which resulted in a drastic confidence drop to 0.1. The importance of MGAT4B decreased to 0.12. The results obtained from this experiment suggest that the information flow through MGAT4B and MGAT5 is crucial for the classification of the analysed patient. MGAT5 as a single marker is not sufficient.

In the last manual counterfactual step, the edge between MGAT5B and MGAT4B was deleted, which finally resulted in a counterfactual, where the label switched from 1 (kidney cancer-specific) to 0 (not kidney cancer-specific). In summary, the conducted experiment demonstrates that human interactions can aid to uncover the most important and relevant parts of a patient graph when guided by recomputed relevance scores after each human counterfactual interaction.

In order to investigate the system itself, as well as its usability, a user study was constructed. The study is created as an online survey that is composed of different tasks and multiple scales to assess the system. This study uses a within-subject design, wherein each participant is presented with all conditions (all three tasks) (Baxter and Jack, 2008). Task instructions are presented to the participant individually for each task. After completing a task the participant is asked to answer a few questions about the interface, such as a change of the confusion matrix, or their thoughts and beliefs of the different actions. If there are any deviations between what happened in the interface and what the participant expected to happen, it can be observed by checking the answers to these questions. Moreover, it can be investigated what caused this confusion e.g. it might be the result of a different expert level or skill set of the participant. To capture more detailed information that might have an impact on the usage, background information about the user is gathered via a short questionnaire at the beginning. Afterwards, tasks with increasing difficulty are presented to the user. Since the usefulness of the interface and the corresponding explanations are to be assessed, questionnaires are added to the end of the study. The System Usability Scale (Brooke, 1996) is used to investigate the usefulness of the interface, and the System Causability Scale (Holzinger et al., 2020) is used to investigate the usefulness of the explanations. Both scales can be rated on a 6-point Likert scale. This is due to the recommendation by Nemoto and Beglar (2014), where it is stated that more detailed results can be obtained by not allowing any “neutral” answers. Hence, the statements are rated on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree).

The study consists of only three tasks, to keep it as short as possible. However, it could be imaginable to test the same primary task, while the participant needs to change the XAI method. This would allow further insight into the understandability and interpretability of each implemented XAI method.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

6 Results

6.1 Part A: Explanation Strategies - impact of Standardization and regulatory activities

RQ: How can we synthesise a common vocabulary concerning XAI and prevent a “tower of Babel effect”, i.e. a confusion of languages?

6.1.1 Standardization efforts concerning XAI

We mapped the ongoing standardization efforts concerning XAI. Standards are developed by (private) organisations, so-called SDOs (Standards Development Organizations). In this field they define concrete technical methods for implementing high-level goals prescribed by law (e.g. transparency). Several SDOs are in the process of developing relevant AI standards concerning XAI. In contrast to the scientific perspective on XAI standardization requires a uniform definition for central terms like explanation as different definitions cannot co-exist and would undermine the goal of standardization.

The joint technical committee JTC 1/SC 42, created by ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission), is drafting ISO/IEC AWI 12792, which aims to create a transparency taxonomy “describing ‘the semantics of the information elements and their relevance to the various objectives of different AI stakeholders”.

The more technically oriented ISO/IEC AWI TS 6254 “Objectives and approaches for explainability of ML models and AI systems” aims at describing “approaches and methods that can be used to achieve explainability objectives of stakeholders with regards to ML models and AI systems‘ behaviours, outputs, and results”. It identifies characteristics of explainability (explanation needs, form, approaches, and technical constraints) and uses them to categorise existing approaches. As a limitation it does not discuss or compare the technological maturity and known limitations of the methodologies (Soler Garrido et al., 2023).

The terms explainability and/or interpretability are also mentioned and defined in ISO/IEC 22989:2022 “Artificial intelligence concepts and terminology” and in ISO/IEC AWI TS 29119-11 concerning the testing of AI systems and in the ISTQB (International Software Testing Qualifications Board) syllabus for “Certified Tester AI Testing” (DIN and DKE, 2022).

The IEEE (Institute of Electrical and Electronics Engineers) is working on the P7000 series of standards as part of the Global Initiative on Ethics of Autonomous and Intelligent Systems. Regarding XAI, the already published standard IEEE P7001 sets out transparency requirements without defining how to achieve them, i.e. which XAI techniques or solutions to use (Soler Garrido et al., 2023).

At the national level the German DIN (Deutsches Institut für Normung) and DKE (Deutsche Kommission für Elektrotechnik Elektronik Informationstechnik) have mapped existing standards and analysed the gaps in standardization as part of a “Standardization Roadmap AI”. Concerning XAI they state that there is a need to specify formal requirements for XAI methods (i.e. formulation of concrete operationalizable/testable requirements) and that additional basic research in XAI is required because available methods have not yet been fully and widely researched and applied (DIN and DKE, 2022).

This mapping shows that at the moment there is still a lack of concrete XAI standards and that standards in development mostly aim at defining central terms and listing transparency desiderata

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

without giving guidance in choosing concrete XAI methods or considering the technical state-of-the-art. This is an important aspect we address with our guidelines presented in part B.

6.1.2 XAI and law

We then mapped the interaction between law and XAI. When processing personal data (e.g. health data as a special category of personal data) in the context of (fully) automated individual decision-making, i.e. without (substantial) human involvement, the General Data Protection Regulation (GDPR) contains duties to inform (Art. 13, 14) and a right to access information (Art. 15) about the 1) "existence of automated decision-making", about 2) "the logic involved" and 3) "the significance and the envisaged consequences of such processing".

The phrase "the logic involved" has generated a version of the "Tower of Babel" effect in legal scholarship. Different interpretations have been proposed, e.g. the "logic involved" necessitates a subject specific local explanation (Malgieri and Comandé, 2017), (Hacker and Passoth, 2022), (Selbst and Powles, 2017) vs a form of general explanation (Wachter et al., 2017) (mainly concerning the features employed on an aggregated level).

A recent opinion (16 March 2023, C-634/21, ECLI:EU:C:2023:220) of the attorney general Pikamäe could clarify the interpretation. These opinions are often but not always adopted by the European Court of Justice. It suggests that a local explanation using XAI is not necessary as only "general information, in particular on the factors taken into account in the decision-making process and their weighting at an aggregated level", i.e. a form of a global feature-importance explanation, has to be provided. But as the opinion also states that "sufficiently detailed explanations on the method used to calculate the score and on the reasons that led to a certain result" have to be provided this seems contradictory to the second statement as the wording "a certain result" seems to imply a local explanation. This contradiction will have to be clarified by the Court of Justice but it seems more likely that "logic involved" will be interpreted as a more general (global) explanation.

Recital 71 also mentions a right "to obtain an explanation of the decision reached after such assessment" as part of suitable measures to safeguard the data subject (Art. 22 para. 3) but this right is only mentioned in the recital. Recitals function as a guide on how to interpret law but cannot create law themselves. Therefore, the existence and the content of a "right to (an) explanation" is still disputed in scholarship (e.g. (Malgieri, 2022). Further clarification by the European Court of Justice is still needed.

In April 2021, the European Commission proposed the so-called Artificial Intelligence Act (AIA). On the 14th of June 2023 the final phase of the law-making process, the so-called Trilogue, began. The AIA follows a risk-based approach, most of its obligations concern so-called high-risk AI systems. Medical devices, which include software (e.g. a diagnosis prediction), are in most cases automatically classified as high-risk AI systems.

The AIA contains AI specific rules for transparency/interpretability for these high-risk AI systems. According to Art. 13 para. 1 (transparency) high-risk AI systems must "be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately". At the moment the necessary level of transparency/interpretability is not sufficiently defined. The AIA does not define the terms "sufficiently transparent" or "to interpret" nor does it mention the concept of "explainability", leading to significant legal uncertainty (Ebers et al., 2021), (Ebers, 2021), (Bomhard & Merkle, 2021). From the viewpoint of the AI developer, it remains unclear if XAI approaches have to be implemented and if a global or local explanation is necessary.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Art. 4a AIA, proposed by the EU Parliament, on general principles defines “transparency” as “AI systems shall be developed and used in a way that allows appropriate traceability and explainability [...] as well as duly informing users of the capabilities and limitations of that AI system and affected persons about their rights” but again does not define “traceability” or “explainability”, leaving question about concrete XAI implementations open to debate and leading to legal insecurity.

It seems more likely that Art. 13 para. 1 AIA does not imply the necessity of explainability in the sense that the way in which data have been processed must be entirely traceable, but a more general form of transparency of the system’s functioning and output generation (Bordt et al., 2022). A recent study commissioned by the European Commission stated that XAI techniques are not the “only means available to understand and interpret AI systems outputs” and therefore not required for all high-risk AI systems. Instead “documentation approaches, scenarios, principles of operations, as well as interactive training materials” will fulfil the requirements (Soler Garrido et al., 2023).

As Art. 13 only concerns the (professional) user (e.g. a doctor), but not a person affected by the AI decision (e.g. a patient), the European Parliament proposed the introduction of “A right to explanation of individual decision-making” (Art. 68c AIA). This would give “any affected person subject to a decision which is taken [...] on the basis of the output from an high-risk AI system” (e.g. a diagnosis by a doctor aided by a model) “which produces legal effects or similarly significantly affects him or her” (e.g. it affects the health of a patient) a “right to request from the deployer clear and meaningful explanation [...] on the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the related input data.” This suggests a form of a local feature-importance explanation (main parameters of the decision, related input data), which could necessitate the implementation of XAI approaches.

Thematically linked, Art. 14 AIA on human oversight also requires the implementation of measures that enable the individuals, to whom human oversight is assigned, to “be able to correctly interpret the high-risk AI system’s output”. In this regard, “the characteristics of the system and the interpretation tools and methods available”, i.e. the implementation of XAI-techniques, have to be taken into account. If XAI approaches become the (technical) state of the art (“tools and methods available”) and if effective human oversight can only be ensured by the aid of XAI approaches this could lead to a duty to implement XAI techniques.

As our mapping of regulatory efforts has shown, there is a high level of legal uncertainty in interpreting these obligations of the AIA. This leads to economic risk for AI providers, who have to interpret the provision themselves. The jurisprudence of the European Court of Justice could lead to clarification, but this will take years.

The ongoing efforts in standardization discussed above are strongly linked with the AIA. AI systems which are in conformity with so-called harmonised-standards, which are developed on demand of the European Commission shall be presumed to be in conformity with the requirements of the AIA. Therefore instead of interpreting the provisions of the AIA themselves, AI providers can mitigate economic uncertainty by following harmonised standards. Therefore, “standards are set to bring the necessary level of technical detail into the essential requirements prescribed in the legal text, defining concrete processes, methods and techniques that AI providers can implement in order to comply with their legal obligations” (Soler Garrido et al., 2023).

As this mapping has shown, it remains unclear if the upcoming AIA obligations require the implementation of XAI and if a global or local explanation is required. As stated, the concrete technical implementation could be clarified by standards. This role of standards can also be criticized as it shifts the law-making power to private bodies, which, compared to national or EU legislation, lack options for democratic control and participation while being vulnerable to lobbying efforts (Ebers,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

2022), (Ebers et al., 2021), (Guijarro Santos, 2023), (Veale & Zuiderveen Borgesius, 2021), (Laux et al., 2023). Therefore, global players could aim at "capturing" the standardization process and define essential XAI terms according to their interests.

6.1.3 A proposed solution

As the mapping of standardization and regulatory efforts has shown, at the moment XAI scientists cannot rely on the vague, partially contradictory, and overly numerous definitions given in the legal and standardization discourses. To pose the opposite problem: how can scientists and XAI scholars inform the process of defining standards and law. The paper (Schneeberger et al., 2023) firstly aims at creating sensitivity about the opacity of the standards drafting mechanism.

Secondly, we proposed a communal initiative to tackle this problem. In recent years, scientists active in the field of XAI have produced several reviews (e.g., (Cambria et al., 2023), (Cabitza et al., 2023), (Vilone and Longo, 2021), (Islam et al., 2022), (Ding et al., 2022), (Hanif et al., 2021), (Haque et al., 2023), (Kargl et al., 2022)), both systematic and more narrative and exploratory ones, to understand the lexical and definition variety in the field and, in some ways, help reduce the linguistic babel.

Building upon these efforts, we proposed to activate a communal initiative that can lead a set of representative scholars to

1) collect all the major definitions proposed in the highest impact articles or most comprehensive reviews

2) invite all the authors of these articles and registered participants at major conferences in the field to vote about the precision, clarity and comprehensiveness of definitions of concepts such as explanation, explainability, transparency, causability and understandability on opportune ordinal scales

3) aggregate the results with state-of-the art methods (e.g. Cabitza et al., 2017) and

4) return the results to the community, possibly iterating a few times so as to reduce variability and facilitate consensus building, in a manner not unlike a Delphi method involving the most motivated people in the field and mediated by asynchronous collaboration tools such as online questionnaires (Shinners et al., 2021) and shared papers.

This lexicographic and definitional effort aims at bringing order and to gaining the necessary visibility and credibility to inform standard and policy making. We propose to start this consolidation process at the Cross Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE) conference 2023, at which the paper (Schneeberger et al., 2023) will be presented. This effort aims at systematically closing the gap between scientific publications and standards as well as regulation.

6.2 Part B: Post-Hoc vs. Ante-Hoc Explanation Strategies - Experimental Evaluation

RQ: How can guidance be provided to non-expert data scientists when choosing between a plethora of different XAI methods?

6.2.1 Interview results

The main interview results of the interviews described in section 4.2.2. above are:

- significant amount of manual work and time required to accurately identify breaking points. This process proved to be labour-intensive and time-consuming, especially for the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

counterfactuals. Concerning counterfactuals and anchors the need for time to carefully select and examine examples was mentioned.

- poor default presentation of almost all techniques. This lack of attention to design made it challenging for users to comprehend and utilize these methods effectively (e.g. the absence of a legend for decision tree components; dissatisfaction with SHAP and BRL graphs). This emphasizes the importance of improving the visual presentation of information to enhance user experience, especially for the default presentation.

- most of the methods were primarily suited for data scientists already familiar with these techniques. This restricted their accessibility to a broader audience, limiting their potential impact.

- all methods required clear and comprehensive explanations themselves. This made them suitable for data scientists who possess the necessary knowledge and expertise to interpret and utilize the provided information effectively, but also widely unusable for nontechnical experts or regular users. All techniques were rated as either ”Not appropriate at all” or ”inappropriate” for regular users by the interviewees, with only the decision trees and anchors being deemed somewhat suitable (”little appropriate” or ”somewhat appropriate”) for non-technical experts.

- concerns regarding the scalability of these methods were mentioned. When dealing with complex models containing numerous features, the interpretability of the results would be challenging.

- Anchors received overall positive feedback as they allowed for the examination of rules applicable to individual samples, making it easier for non-technical experts to understand. Anchors were appreciated for their ability to present both general rules and rules specifically applicable to each sample.

- Counterfactuals were deemed useful for establishing general rules and ensuring comprehensive coverage of the data

- Generalized Additive Models (GAMs) were commended for their visually appealing nature and their ability to capture nonlinearity effectively

- BRL was mentioned as appearing easy at first glance but lacking sufficient insight upon closer inspection.

- concerning the desired features for XAI guidelines concrete numbers for bar plots in SHAP were suggested to provide clearer visual representations of the data. Additionally, concerns were raised about the classification of samples in SHAP and the clarity of horizontal bars, which needed improvement.

- it is crucial for explanations to be concise and understandable within a short timeframe.

- explanations must be accessible even to individuals with limited technical knowledge or expertise

- different opinions about the maximum acceptable waiting time for obtaining explanations (e.g. one day for complex models vs. the need for real-time explanations)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6.2.2 XAI decision tree

As a key outcome, we created a decision tree (Figure 3) that serves as a valuable aid for selecting the most suitable xAI technique based on specific considerations such as target users, performance, fidelity, completeness, and performance requirements. The purpose of this tool is twofold: first, to aid readers in comprehending the trade-offs associated with different xAI methods; and second, to illustrate how these methods can be combined synergistically, resulting in a comprehensive understanding of ML models by employing the most appropriate tools for the task at hand.

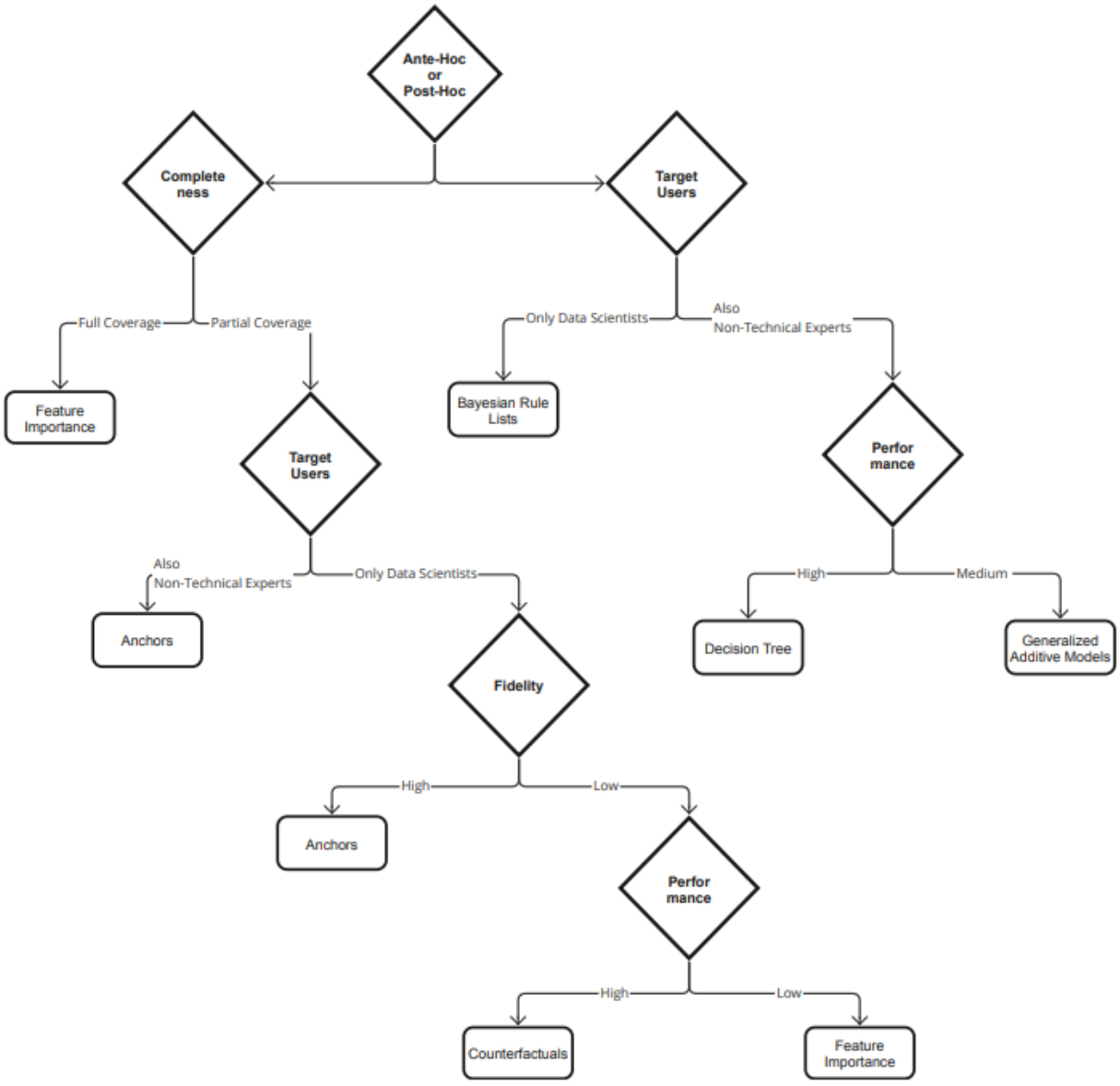

Figure 3: Decision Tree for the usage of XAI methods (Retzlaff et al., 2023).

The decision process for determining the appropriate explainability method for AI and ML approaches involves several key considerations. First, the choice between post-hoc and ante-hoc methods depends on specific requirements regarding the complexity and scale of neural networks and subsequent models. If performance is the primary concern, the use of neural networks can be requisite, and with that require the use of Post-hoc methods. However, if either the data availability of computational requirements is not met or model simplicity and understandability are also a

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

concern, ante-hoc models may be the better choice, as they offer simplicity and ease of use in addition to performance considerations.

Within the ante-hoc methods, the first decision is made based on the target users. Bayesian rule lists (BRL) are suitable for data scientists, while decision trees and generalized additive models (GAMs) can also cater to non-technical experts, providing a wider range of usability.

The next decision then involves evaluating whether faster generation time and inference, offered by decision trees, are necessary or if GAMs can also meet the requirements. Since the performance difference between them is existent but not substantial, other factors such as data scaling and multicollinearity should be taken into account during the decision-making process.

When considering post-hoc methods, the first criterion is completeness. If full coverage of the model is required, feature importance scores should be used. However, if the focus is on local explanations, anchors and counterfactuals can also be considered.

The next decision point revolves around the target users. If the aim is to communicate the results to non-technical experts, anchors should be prioritised due to their ability to provide easily understandable insights. Conversely, if the communication is limited to data scientists, all post-hoc methods are deemed suitable based on the evaluation.

Fidelity is the next aspect to consider. If high fidelity of individual explanations is necessary, anchors should be chosen. This approach stands out as the only one capable of delivering truthful anchoring of the provided examples. In cases where absolute fidelity is not required, counterfactuals and feature importance can also be taken into account.

The subsequent decision node revolves around performance, where counterfactuals offer significantly faster computation compared to feature importance. When explanations need to be generated on large datasets with numerous features or on low-end devices within time-constrained contexts, counterfactuals should be preferred over feature importance approaches.

6.3 Part C: Usability Evaluation CLARUS

RQ: Is the interface useful to manipulate PPI networks, and if so, to what extent? How easy is it for the “human-in-the-loop” to manipulate tasks in the XAI-platform?

Domain experts were recruited to conduct the study. This was due to the specific use case of the interface. The study demonstrates the benefit of CLARUS in terms of knowledge gain and causal understanding of the user, for instance on predictions and their explanations. In total, 31 participants finished the survey.

In the beginning, the participants were asked to provide some background knowledge about their occupation and expertise level/familiarity with the topics of Neural Networks (NN), Protein-Protein Interaction (PPI) graphs and explainable AI (XAI). Results from this query are shown below.

The occupations of the participants are listed below. They indicate a fair share of Computer and Mathematical Scientists and Life, Physical, and Social Science Occupations.

- Computer and Mathematical Occupations 14
- Life, Physical, and Social Science Occupations 11
- Education, Training, and Library Occupations 3
- Business and Financial Operations Occupations 1
- Healthcare Occupations 1

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

The mean age of the participants is ~36 years, whereby the youngest participant is 22 and the oldest 54 years old. The distribution of the participant's age is shown in Figure 4.

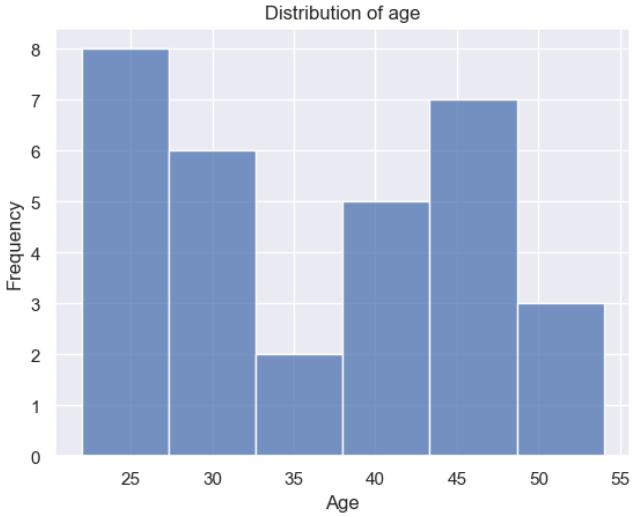

Figure 4: Age distribution of participants.

The rating for the background knowledge of each participant can be observed in Figure 5.
The mean familiarity with each of the topics is:

- Familiarity with Neural Networks (NNs): 3.97
- Familiarity with Protein-Protein-Interaction (PPI) Networks: 2.55
- Familiarity with explainable AI (XAI): 3.39

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

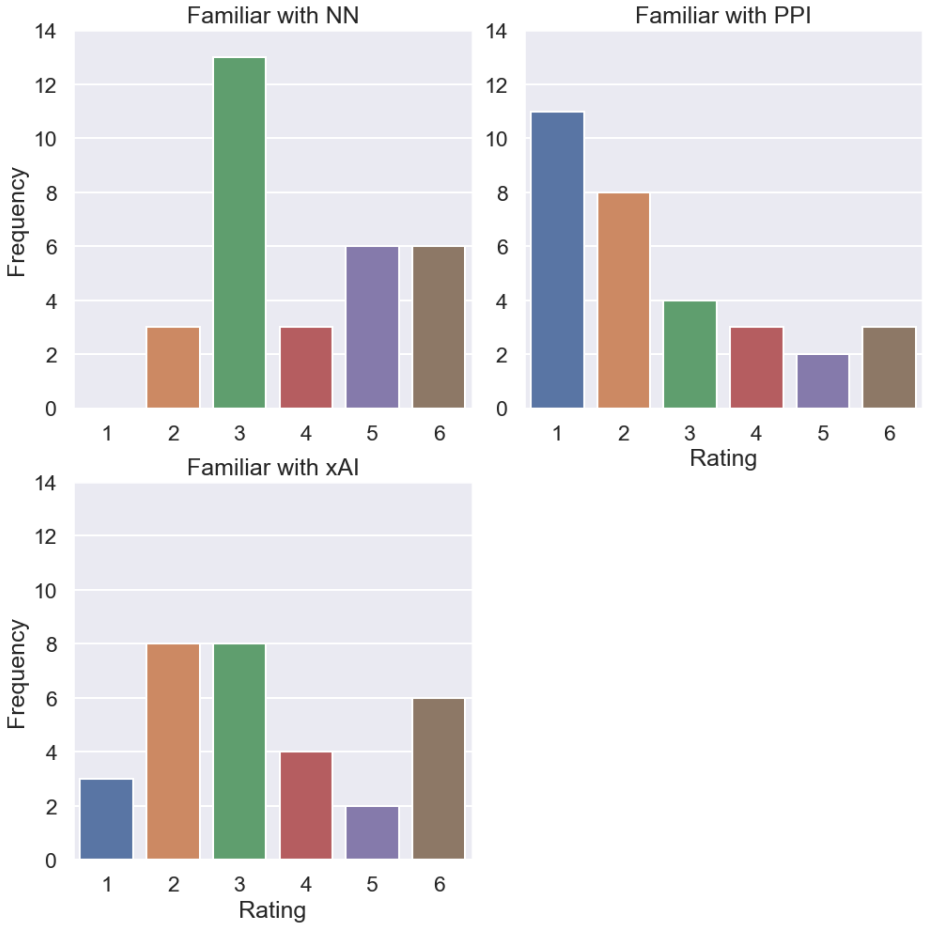

Figure 5: Participant ratings per Background question.

6.3.1 Target user group

Afterwards, the participants were asked to visit the interface (http://rshiny.gwdg.de/apps/clarus/) and familiarise themselves with it. After interacting with the interface, participants were asked to state their beliefs about the target user group.

Most participants stated that they think the target user group for this application are Biological experts, Biologists/Scientists, and medical researchers. Others also mentioned Bioinformaticians, Data Scientists, Computational biologists, omics researchers and anyone investigating Graph Neural Networks (GNNs).

For example, one user stated, *"The target user group will definitely be biologists or biological researchers - for me as a medical expert it seems to be very hard ..."*.

6.3.2 Tasks - Interpretation and understanding of functions/metrics

Then the participants were asked to perform different tasks, such as deleting a node and investigating the changes, stating their expectations and the actual behaviour of the interface. To investigate the difference between user expectations, their interpretations, the understanding of the functionality and the behaviour of the system, three tasks were constructed.

First, participants were asked to delete the node with the highest degree, forcing them to investigate the sorting or highlighting of the nodes according to their degree. This task also investigated the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

understanding of the confusion matrix and the expected changes for such an action. The format for all tasks was similar, as an example, Figure 6 shows a screenshot of the first task and the corresponding questions. The following results were obtained from the first task.

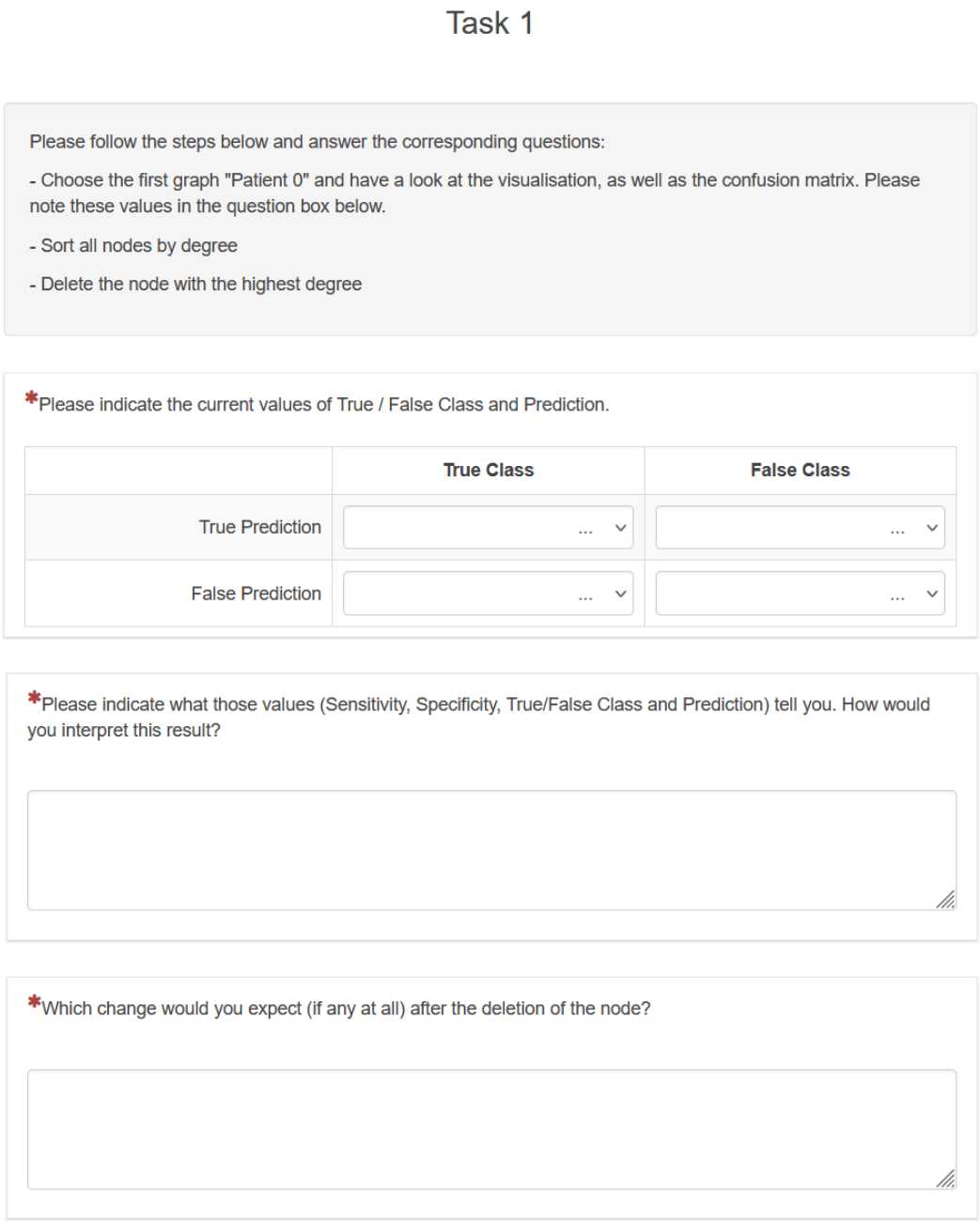

Figure 6: Survey screenshot of the first task.

Participant understanding of the provided confusion matrix and its values was quite high, meaning most participants did not face any issues finding and interpreting the values of the confusion matrix. The exact number of participants and their answers are summarised below. Hereby, “Not visible” means that the participant was not able to view or find the confusion matrix, “No answer/Server crash” indicates that the participant did not state a useful answer, or that the server crashed and thus no answer was given. Further, one participant was unsure about the values of the confusion matrix, whereas the other 25 participants were able to explain and interpret the presented values.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

- Not visible: 1
- No answer / Server crash: 3
- Participants are unsure: 1
- Participants did understand the values (have given textual answers): 26

The second task asked the participants to delete one node with the most edges and perform the "predict" action. Once again, participants were asked to indicate their expectations for the corresponding action and the perceived changes in the interface. The following results were obtained from the second task.

Most participants were able to give an explanation or state their beliefs about the functionality of "predict". Thus, the majority of participants did understand the function and its results. They stated their expectations and interpretations of the functionality. Again, 2 participants reported server issues or gave a textual answer such as "Na". The category "Not clear" indicates that the participants did not know what the "predict" functionality is supposed to do, or how to interpret the results.

- No answer / Server crash: 2
- Nothing/ unclear until testing it: 6
- New estimate / change in confusion matrix / no change expected: 23

In the third and last task participants were asked to add a new node, with given parameters, and create a new edge between the new node and one with the current highest degree. Then they were asked to perform the "retrain" action. Again, questions regarding their expectations, understanding of the action and the actual behaviour were raised. The results were obtained as follows.

It was observed that fewer participants than before were able to understand and interpret the functionality of "retrain". Some participants expected something different from the already viewed functions, those are summarised in the "Not visible" category since they stated that they were not able to observe the function. One participant stated that the node added before was not visible anymore, this is summarised as "other issues". Other participants either had no expectations regarding the functionality, or it was unclear to them.

- Unclear / no expectations: 8
- Server crash / other issues: 3
- Nothing happened / unclear: 2
- Change in metrics/ subnet shape / no change expected: 18

6.3.3 Usability and Causability

After finishing these tasks, participants were asked to rate the usability and causability/interpretability according to the System Usability Scale (SUS) (Brooke, 1996) and the System Causability Scale (SCS) (Holzinger et al., 2020) respectively. The questions of the SUS are alternating positive and negative statements, to prevent a response bias (Brooke, 2013).

As mentioned before, 6-point Likert scales were used for the ratings. The score is calculated by the sum of the score contributions from each item (Holzinger et al., 2020). In this case, the score contributions range from 0 to 5. Because of the introduced alteration of the questions, the score of odd-numbered questions is computed by the scale position minus one. For all even question numbers, the contribution is computed by 6 minus the scale position. The sum is then multiplied by two, thus reaching a possible maximum of 100 points. According to Bangor, Kortum and Miller (2009), a score below 50 points is not acceptable regarding usability. A score above 68 is considered above average, whereas a score below 68 is considered below average (Lewis et al., 2018).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

The score was calculated for each participant, ranging from the worst assessment of 10 to the best of 100. The mean value of the Usability score is **57.61**. This shows very clearly how subjective the rating of usability really is. However, on average the system is not too bad, considering its innovative concept and needed level of expertise. Nevertheless, this shows that the prototype is useful and can be improved to better suit the needs of the experts.

Since the SCS follows a similar pattern, the scale is computed in the same fashion. The mean score of Causability, measured for all participants is **52.0**. Whereby 40 is the lowest, and 62 is the highest score respectively. This shows that the explanations were understandable and thus helped the experts to interpret the changes of the predictions. However, this value is also an indicator of the improvement potential of the interface.

6.3.4 Feedback

At the end of the questionnaire, participants were able to provide feedback regarding the interface. The main take-away points and important comments are summarised.

Most participants reported an extremely slow interaction with the interface. Some also experienced server crashes. This could be due to an overload of the server's network traffic. However, we also noted some issues in the code, that could lead to a crash or long waiting times.

Interestingly, a main message from multiple participants was that the interface itself and the functionality were unclear. This means that people are able to use the interface and execute given instructions, however, they would not be able to use the system efficiently by themselves.

Key aspects of the interface are the explanations and their opportunity to allow experts a better insight into AI-generated predictions. According to the Causability ratings, the system provides this functionality to some extent. Though, most participants seemed to struggle with those explanations. Some even stated in the feedback that they are unsure about the explanations, where to find them and if there were any at all. Participants wished for further information, hints or even a demo on how to find, use and interpret the explanations.

Something that should also be noted is, according to the participants, the general lack of hints or descriptions. Starting from the dataset, where some participants would have liked more information on the features of each node, to get a better overview of the data. It was also stated that the information on the general process of the interface was missing. Thus, participants did not really know why they interacted with the system, or which benefit this might bring them, but merely executed the given tasks. Here again, participants noted that more hints or a complete demo on how to use the interface and interpret the explanations and changes thereof would be necessary. Further, it could also be observed for the tasks, that some participants did not know what to expect from the functions "predict" and "retrain". Thus, there should be a description of some sort, to help users understand the core functionality of the system.

Other usability-related issues were raised regarding the feedback of the system state. Participants expected some kind of notification, when the dataset finished loading, or parameters were saved for a new node. They found it confusing that a list field was used for selecting one feature at a time, to enter the corresponding value of a new node. This could be improved by presenting multiple input fields, one for each feature. However, if a network assigns a hundred features to a node, this could be extremely convoluted. Short keys or keyboard interaction was expected from one participant.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Suggested Improvements

To increase the usability and Causability of the system, participants suggested improvements. The most important ones are noted and summarised.

1. Better / Faster Server
2. Highlight / describe the explanations and their meaning
3. Add demos and hints for the functionalities
4. Better structure of the overall system & more feedback
    a. Feedback when loading is finished
    b. Adding a node should be refactored -> show both/all features at once
    c. Maybe add the dataset name to the interaction tab
    d. Participant recommendation: interactions left to the graph; confusion matrix on top; toggle values inside the graph; align both tables; use i-Icon for hints

Since there were some issues with the server, participants recommended using a different, or faster server. This does not directly link to the results of the study, however, a stable connection for a multi-user system should not be neglected to provide at least the minimal requirements to be usable.

A point of utmost importance, especially considering the use case of the application, is that there are nearly no hints or explanations for an inexperienced user. Most participants recommended adding descriptions and hints to each functionality and thus providing a better and easier start. Although the Causability scale was not too bad, participants felt the need for additional information and explanations on how to interpret certain values or actions. This also sparked the need for some kind of demo. Users might not feel as overwhelmed if there was detailed documentation provided on how to use the interface and interpret the explanations. Demos may benefit not only inexperienced users or non-domain experts but also provide guidance to domain experts or support experienced users.

The evaluation of usability showed another critical factor, system feedback. Feedback on the current system state to the user is considered essential for the usability of a system (Nielsen and Molich, 1990). Thus, it represents one of the ten usability heuristics by Nielsen and Molich (1990). The interface should notify users when data is finished loading, or while computations are executed. Moreover, the structure of the system and the corresponding placement of components seemed confusing to some participants. Thus, a refactoring of the prototype design and some methods was proposed.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

7 Conclusion

Our approach to the three different parts (A, B, C) made it possible to dive deeper into the topic of XAI and allow different views from various scientific disciplines. In part A we were able to present a map of the ongoing standardisation efforts, which aim to define foundational XAI terms like transparency or explainability and standardise the concrete technical implementation of XAI techniques. We analysed the interplay of XAI and law with a special focus on the proposed Artificial Intelligence Act. As a result of those analyses of standardisation and law, we proposed a communal initiative, which attempts to bring order into the confusion of languages concerning XAI. Part B highlights the importance of explanations and which XAI method is best suitable for a specific purpose. Moreover, the resulting article serves as a practical guideline for data scientists and AI engineers, when planning to deploy an AI model. These guidelines promote the use of XAI methods while keeping the decision criteria rather simple. Hence, they help data scientists or programmers to decide on the best suitable XAI method and incorporate them into their systems. This is also in line with the future strategy of the European Union since transparency and re-traceability are promoted through these methods. In Part C, we evaluated our own approach to an interactive XAI platform for medical practitioners, scientists and researchers. It was found that the prototype has room for improvement, however, it was deemed useful by the survey participants. Thus, it could be observed that explanations benefit user understanding while experimenting with complex problems such as protein-protein interaction networks.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

8 References

- 'Artificial Intelligence Act'. in: DiMatteo, L. A., Poncibò, C. & Cannarsa, M. (eds.), The Cambridge Handbook of Artificial Intelligence: Global Perspectives on Law and Ethics. Cambridge University Press, 321-344, doi: 10.1017/9781009072168.030.
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115, doi: 10.1016/j.inffus.2019.12.012.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one, 10, (7), e0130140, doi: 10.1371/journal.pone.0130140.
- Bangor, A., Kortum, P., and Miller, J. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. Journal of Usability Studies, 4, (3), pp. 114–123.
- Baxter, P., & Jack, S. (2008). Qualitative case study methodology: Study design and implementation for novice researchers. The Qualitative Report, 13, (4), pp. 544-559.
- Beinecke, J., Saranti, A., Angerschmid, A., Pfeifer, B., Klemt, V., Holzinger, A. & Hauschild, A.-C. (2022). CLARUS: An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks. bioRxiv, 2022.11. 21.517358, doi:10.1101/2022.11.21.517358.
- Bevan, N. (1995). Measuring Usability as Quality of Use. Software Quality Journal, 4, (2), 115--130, doi: 10.1007/BF00402715.
- Bomhard, D. & Merkle, M. (2021). Europäische KI-Verordnung. Recht Digital 1(6), 276-283.
- Bordt, S., Finck, M., Raidl, E. & von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. in: FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, ACM, 891-905. ACM, doi: 10.1145/3531146.3533153.
- Brooke, J. (1996). SUS-A quick and dirty usability scale, in P. W. Jordan, B. Thomas, Ian Lyall McClelland, Bernard Weerdmeester (eds.), Usability evaluation in industry, 1st edn, CRC Press, London, pp. 4-7.
- Brooke, J. (2013). SUS: A Retrospective. Journal of User Experience, 8, (2), pp. 29-40.
- Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K. & Holzinger, A. (2023). Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI. Expert Systems with Applications, 213, (3), 118888, doi:10.1016/j.eswa.2022.118888.
- Cabitza, F., Ciucci, D. & Locoro, A. (2017). Exploiting collective knowledge with three-way decision theory: cases from the questionnaire-based research. International journal of approximate reasoning 83, 356-370, doi: 10.1016/j.ijar.2016.11.013.
- Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M. & Nobani, N. (2023). A survey on XAI and natural language explanations. Information Processing & Management 60(1), doi: 10.1016/j.ipm.2022.103111.
- Castelvecchi, D. (2016). Can we open the black box of AI? Nature News, 538, (7623), 20-23, doi:10.1038/538020a.
- Dandl, S., Molnar, C., Binder, M., Bischl, B. (2020). Multiobjective counterfactual explanations, in: Bäck, T., Preuss, M., Deutz, A., Wang, H., Doerr, C., Emmerich, M.,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Trautmann, H. (eds.), Parallel Problem Solving from Nature – PPSN XVI, Springer International Publishing, 448-469. doi:10.1007/978-3-030-58112-1_31.

- DIN & DKE. (2022). Normungsroadmap Künstliche Intelligenz: Version 2. https://www.dke.de/de/arbeitsfelder/core-safety/normungsroadmap-ki.
- Ding, W., Abdel-Basset, M., Hawash, H. & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. Information Sciences 615, 238-292, doi: 10.1016/j.ins.2022.10.013.
- Ebers, M. (2021). Standardisierung Künstlicher Intelligenz und KI-Verordnungsvorschlag. Recht Digital 2, 588-597.
- Ebers, M. (2022). Standardizing AI: The case of the European Commission's proposal for an
- Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H. & Steinrötter, B. (2021). The European Commission's Proposal for an Artificial Intelligence Act: A Critical Assessment by Members of the Robotics and AI Law Society (RAILS). J 4(4), 589-603, doi: 10.3390/j4040043.
- Goldstein, A., Kapelner, A., Bleich, J. & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics, 24, (1), 44-65, doi:10.1080/10618600.2014.907095.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F. & Pedreschi, D. (2019). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51, (5), 93, doi:10.1145/3236009.
- Guijarro Santos, V. (2023). Nicht besser als nichts: Ein Kommentar zum KI-Verordnungsentwurf.
- Hacker, P. & Passoth, J.H. (2022): Varieties of AI explanations under the law: From the GDPR to the AIA, and beyond. in: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R. & Samek, W. (eds.), xxAI: Beyond Explainable AI, Springer International, 343-373, doi: 10.1007/978-3-031-04083-2.
- Hanif, A., Zhang, X. & Wood, S. (2021). A survey on explainable artificial intelligence techniques and challenges. in: 2021 IEEE 25th international enterprise distributed object computing workshop (EDOCW), IEEE, 81-89, doi: 10.1109/EDOCW52865.2021.00036.
- Haque, A. B., Islam, A. N. & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technological Forecasting and Social Change 186, doi: 10.1016/j.techfore.2022.122120.
- Hastie, T. J. (1992). Generalized Additive Models, in: Chambers, J. M. & Hastie, T.J. (eds.), Statistical Models in S, Routledge, 249-308.
- Holzinger, A. & Mueller, H. (2021). Toward Human-AI Interfaces to Support Explainability and Causability in Medical AI. IEEE COMPUTER, 54, (10), 78-86, doi:10.1109/MC.2021.3092610.
- Holzinger, A. (2005). Usability engineering methods for software developers. Communications of the ACM, 48, (1), 71-74, doi: 10.1145/1039539.1039541.
- Holzinger, A. (2016). Interactive Machine Learning for Health Informatics: When do we need the human-in-the-loop? Brain Informatics, 3, (2), 119-131, doi:10.1007/s40708-016-0042-6.
- Holzinger, A. (2020). Explainable AI and Multi-Modal Causability in Medicine. Wiley i-com Journal of Interactive Media, 19, (3), 171--179, doi:10.1515/icom-2020-0024.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Holzinger, A., Carrington, A. & Mueller, H. (2020). Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. KI - Künstliche Intelligenz, 34, (2), 193--198, doi:10.1007/s13218-020-00636-z.
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K. & Müller, H. (2019a). Causability and Explainability of Artificial Intelligence in Medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, (4), 1-13, doi:10.1002/widm.1312.
- Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G.C., Pintea, C.-M. & Palade, V. (2019b). Interactive machine learning: experimental evidence for the human in the algorithmic loop. Applied Intelligence, 49, (7), 2401-2414, doi:10.1007/s10489-018-1361-5.
- Holzinger, A., Saranti, A., Molnar, C., Biecek, P. & Samek, W. (2022). Explainable AI Methods - A Brief Overview. XXAI - Lecture Notes in Artificial Intelligence LNAI 13200. Cham: Springer, pp. 13--38, doi:10.1007/978-3-031-04083-2_2.
- Islam, M. R., Ahmed, M. U., Barua, S. & Begum, S. (2022). A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Applied Sciences 12(3), doi: 10.3390/app12031353.
- Kargl, M., Plass, M. & Müller, H. (2022). A literature review on ethics for AI in biomedical research and biobanking. Yearbook of Medical Informatics 31(1), 152–160, doi: 10.1055/s-0042-1742516.
- Lakkaraju, H., Kamar, E., Caruana, R. & Leskovec, J. (2017). Interpretable and Explorable Approximations of Black Box Models. arXiv:1707.01154.
- Lakkaraju, H., Kamar, E., Caruana, R. & Leskovec, J. Faithful and customizable explanations of black box models. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), (2019).
- Laux, J., Wachter, S. & Mittelstadt, B. (2023). Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act, preprint https://papers.ssrn.com/sol3/papers.cfm?abstractid = 4365079.
- Letham, B., Rudin, C., McCormick, T. H., Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model, The Annals of Applied Statistics, 9(3), 1350-1371, doi:10.1214/15-AOAS848.
- Lewis, James R., & Sauro, J. (2018). Item benchmarks for the system usability scale. Journal of Usability Studies, 13, (3), pp. 158-167.
- Lundberg, S. M. & Lee, S.-I. (2017). 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, Curran Associates, Inc., 4765-4774.
- Malgieri, G. & Comandé, G. (2017). Why a right to legibility of automated decision-making exists in the general data protection regulation. International Data Privacy Law 7(4), 243-265, doi: 10.1093/idpl/ipx019
- Malgieri, G. (2022). Automated decision-making and data protection in Europe. in: González Fuster, G., van Brakel, R. & De Hert, Paul (eds.), Research Handbook on Privacy and Data Protection Law, Edward Elgar, 433-448, doi: 10.4337/9781786438515.
- Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd Edition. Independently published.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. (2019). Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning. Cham: Springer/Nature, pp. 193--209, doi:10.1007/978-3-030-28954-6_10.
- Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires, in Sonda, N. & Krause, A. (eds.), JALT2013: Learning is a Lifelong Voyage, Japan Association for Language Teaching, Taito, Japan, pp. 1-8.
- Nielsen, J. & Molich, R. (1990). Heuristic evaluation of user interfaces, in JC Chew & J Whiteside (eds.), Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '90). Association for Computing Machinery, New York, NY, USA, pp. 249–256. doi:10.1145/97243.97281.
- Pfeifer, B., Saranti, A., & Holzinger, A. (2022). 'GNN-SubNet: Disease subnetwork detection with explainable graph neural networks.', Bioinformatics 38.Supplement_2 (2022): ii120-ii126. doi: 10.1093/bioinformatics/btac478.
- Plass, M., Kargl, M., Evans, T., Brcic, L., Regitnig, P., Geißler, C., Carvalho, R., Jansen, C., Zerbe, N., Holzinger, A. & Müller, H. (2023). Human-AI Interfaces are a Central Component of Trustworthy AI. In: Mehta, Mayuri, Palade, Vasile & Chatterjee, Indranath (eds.) Explainable AI: Foundations, Methodologies and Applications. Cham: Springer International, pp. 225--256, doi:10.1007/978-3-031-12807-3_11.
- Retzlaff, C.O., Angerschmid, A., Saranti, A., Schneeberger, D., Roettger, R., Mueller, H. & Holzinger, A. (2023). Post-Hoc vs Ante-Hoc Explanations: XAI Design Guidelines for Data Scientists. Cognitive Systems Research, in preparation.
- Ribeiro, M. T., Singh, S., Guestrin, C. (2018). Anchors: High precision model-agnostic explanations, Proceedings of the AAAI Conference on Artificial Intelligence 32(1), doi:10.1609/aaai.v32i1.11491.
- Safavian S. & Landgrebe D. (1991). A survey of decision tree classifier methodology, IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660-674, doi:10.1109/21.97458.
- Schneeberger, D., Roettger, R., Cabitza, F., Campagner, A., Plass, M., Mueller, H. & Holzinger, A. (2023). The Tower of Babel in explainable Artificial Intelligence (XAI). Springer Lecture Notes in Computer Science (LNCS) Volume 14065. Springer. in print.
- Selbst, A. D. & Powles, J. (2017). Meaningful information and the right to explanation. International Data Privacy Law 7(4), 233-242, doi: 10.1093/idpl/ipx022.
- Shinners, L., Aggar, C., Grace, S. & Smith, S. (2021). Exploring healthcare professionals' perceptions of artificial intelligence: Validating a questionnaire using the e-delphi method. Digital Health 7, doi: 10.1177/20552076211003433.
- Soler Garrido, J., Tolan, S., Hupont Torres, I., Fernandez Llorca, D., Charisi, V., Gomez Gutierrez, E., Junklewitz, H., Hamon, R., Fano Yela, D. & Panigutti, C. (2023). AI Watch: Artificial intelligence Standardisation Landscape Update. https://publications.jrc.ec.europa.eu/repository/handle/JRC131155.
- Veale, M. & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act. Computer Law Review International 22, 97-112, doi: 10.9785/cri-2021-220402.
- Vidovic, M.M.-C., Görnitz, N., Müller, K.-R., Rätsch, G. & Kloft, M. (2015). Opening the Black Box: Revealing Interpretable Sequence Motifs in Kernel-Based Learning Algorithms. Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, pp. 137-153, doi: 10.1007/978-3-319-23525-7_9.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

- Vilone, G. & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion 76, 89-106, doi: 10.1016/j.inffus.2021.05.009.
- Wachter, S., Mittelstadt, B. & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. International Data Privacy Law 7(2), 76-99, doi: 10.1093/idpl/ipx005.
- Ying, Z., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks, in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett (eds.), Advances in Neural Information Processing Systems, 32.
- Zador, A., Escola, S., Richards, B., Ölveczky, B., Bengio, Y., Boahen, K., Botvinick, M., Chklovskii, D., Churchland, A. & Clopath, C. (2023). Catalyzing next-generation artificial intelligence through neuroai. Nature communications, 14, (1), 1597, doi: 10.1038/s41467-023-37180-x.
- Zeitschrift für Digitalisierung und Recht 3(1), 23-42.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.