Does this Explanation Help? Designing Local Model-agnostic Explanation Representations and an Experimental Evaluation using Eye-tracking Technology

MIGUEL ANGEL MEZA MARTÍNEZ, Karlsruhe Institute of Technology (KIT), Germany
MARIO NADJ, TU Dortmund University, Germany
MORITZ LANGNER, Karlsruhe Institute of Technology (KIT), Germany
PEYMAN TOREINI, Karlsruhe Institute of Technology (KIT), Germany
ALEXANDER MAEDCHE, Karlsruhe Institute of Technology (KIT), Germany

In Explainable Artificial Intelligence (XAI) research, various local model-agnostic methods have been proposed to explain individual predictions to users in order to increase the transparency of the underlying Artificial Intelligence (AI) systems. However, the user perspective has received less attention in XAI research, leading to a (1) lack of involvement of users in the design process of local model-agnostic explanations representations and (2) a limited understanding of how users visually attend them. Against this backdrop, we refined representations of local explanations from four well-established model-agnostic XAI methods in an iterative design process with users. Moreover, we evaluated the refined explanation representations in a laboratory experiment using eye-tracking technology as well as self-reports and interviews. Our results show that users do not necessarily prefer simple explanations and that their individual characteristics, such as gender and previous experience with AI systems, strongly influence their preferences. In addition, users find that some explanations are only useful in certain scenarios making the selection of an appropriate explanation highly dependent on context. With our work, we contribute to ongoing research to improve transparency in AI.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: machine learning, explainability, user-centric evaluation, eye-tracking

1 INTRODUCTION

Artificial Intelligence (AI) is playing an increasingly significant role in high-stake domains such as finance [10], criminal justice [40], and healthcare [88]. Nevertheless, the effectiveness of AI systems is limited by their inability to explain their decisions to human users [96]. Specifically, many AI systems are opaque, and their underlying Machine Learning (ML) models are considered "black boxes" where only the model's output is available. This lack of transparency leaves the inner working mechanisms of the AI system unclear to users. This problem is reinforced by the popularity of deep learning models, which are hard to understand even by experts [23]. Due to the need for users to understand the logic of these systems, new regulations have been enacted that provide the

Authors' addresses: Miguel Angel Meza Martínez, miguel.martinez@kit.edu, Karlsruhe Institute of Technology (KIT), P.O. Box 6980, Karlsruhe, Baden-Württemberg, Germany, 76049; Mario Nadj, mario.nadj@tu-dortmund.de, TU Dortmund University, Friedrich-Wöhler-Weg 6, Dortmund, Nordrhein-Westfalen, Germany, 44227; Moritz Langner, moritz.langner@kit.edu, Karlsruhe Institute of Technology (KIT), P.O. Box 6980, Karlsruhe, Baden-Württemberg, Germany, 76049; Peyman Toreini, peyman.toreini@kit.edu, Karlsruhe Institute of Technology (KIT), P.O. Box 6980, Karlsruhe, Baden-Württemberg, Germany, 76049; Alexander Maedche, alexander.maedche@kit.edu, Karlsruhe Institute of Technology (KIT), P.O. Box 6980, Karlsruhe, Baden-Württemberg, Germany, 76049.

2160-6455/2023/7-ART $15.00
https://doi.org/10.1145/3607145

"right to explanations" for all decisions made or supported by AI systems [17, 23]. Nevertheless, such regulations still need to be put into practice. Overall, it remains difficult for non-experts to understand the logic behind AI systems and figure out how specific inputs lead to a particular output [17].

In this light, there has been a surge of interest in Explainable Artificial Intelligence (XAI) among scholars and practitioners seeking to produce models that can explain their decisions in non-technical terms while maintaining a high prediction accuracy [22, 94]. XAI aims to support human understanding and trust in the use of AI systems [94]. In particular, extensive research has focused on developing so-called "model-agnostic" explainable methods. These methods are applied after the predictive model has been trained and can explain any predictive model [87]. Some of these model-agnostic methods provide local explanations, which clarify how individual predictions are made. In contrast, others offer global explanations, which describe the entire model behavior across all instances for a given dataset [68]. This article focuses on model-agnostic methods that provide local explanations (e.g., LIME [81]) as they provide great flexibility by isolating the explanations from the underlying predictive model [80]. As a result, practitioners can assess and compare different models and even migrate to a new model later [2, 68]. Furthermore, users can understand why AI systems make particular decisions [12, 87].

However, despite the advances within the XAI field, challenges remain. First, strong critiques have been made that explanation representations are still based on researchers' intuition rather than a deep understanding of users' needs [65, 96]. Among others, Miller [65] argues that experts who train and evaluate models may lack the judgment necessary to assess the usefulness of the generated explanation representations for users (see also [2, 67, 83]). Thus, it is necessary to involve users in the design process when creating or refining local model-agnostic explanation representations. An appropriately designed explanation representation can help increase users' trust and understanding of the AI system's decisions [13, 23].

Second, to evaluate XAI explanations with users, researchers commonly rely on a combination of objective performance measures on an evaluation task and subjective self-reported measures of constructs such as trust and understandability [12, 24]. Nevertheless, studies have found that objective performance measures might not generalize well from the evaluation task and that subjective measures might not predict the utility of explanations in decision-making tasks [12]. Furthermore, subjective measures only provide a limited view of how users utilize explanations during an evaluation task, as their data collection occurs after task completion [71]. Therefore, researchers have argued for incorporating complementary data collection methods based on non-intrusive measures that provide insight into how users actually utilize explanations during the evaluation task [71]. In particular, there seems to be an increasing interest in incorporating the use of eye-tracking technology in order to collect data on users' visual attention (see, e.g., [8, 31, 71, 85]). The use of eye-tracking technology in research has considerably increased in recent years with the development of more accurate and affordable devices due to their capacity to provide insights into users' cognitive processes and, specifically, visual attention [28, 42, 60, 79]. For instance, eye-tracking has been used to analyze users' comprehension of graphs [1, 89] and to investigate how users evaluate visual analytics [57]. Nevertheless, the use of eye-tracking technology has not received much attention in evaluating local model-agnostic explanation representations of XAI methods.

Against this backdrop, the goal of this article is to address the following research questions:

RQ1*: How to design local model-agnostic explanation representations from different XAI methods following an iterative design process with users?*

RQ2*: How do users perceive, evaluate, and visually attend the designed local model-agnostic explanation representations from different XAI methods?*

To achieve this goal, we first refined representations of local explanations from four well-established model-agnostic XAI methods following an iterative design process involving users [48]. This iterative refinement allowed us to increase the explanation representations' comparability while controlling for any confounding factors due to their different explanation approaches. The selected methods are Anchors [82], Diverse Counterfactuals Explanation (DICE) [69], Local Interpretable Model-agnostic Explanations (LIME) [81], and Shapley Additive

Explanations (SHAP) [62] (a detailed overview of each method and the reason for their selection is provided in Section 3.2). After the iterative refinement of the explanation representations, we evaluated them with 19 participants in a laboratory experiment using eye-tracking technology and self-reports, followed by interviews. We centered our evaluation in the bank loan applications domain, where AI models predict the decision to approve or reject a loan application by evaluating its risk using a set of attributes. Details regarding the selected domain and dataset used are found in Section 3.1.

The contributions of our work are two-fold. First, we performed an iterative design process involving users in refining the representations of local explanations from well-established model-agnostic methods. This iterative design process provides insightful information on how researchers can increase the comparability of explanation representations for well-established XAI methods. Moreover, users' evaluations throughout the iterative design process inform how they perceive the explanation representations. Second, we leverage eye-tracking technology to evaluate how users visually attend to these explanation representations in a laboratory experiment. This approach extends the work from a limited number of XAI studies that integrated eye-tracking technology by evaluating multiple model-agnostic XAI methods with users [9, 18, 19, 52, 70, 75]. As a result, we provide an interesting perspective on how users utilize different explanation representations and which ones they prefer. Additionally, as a practical contribution, we provide an open-source reference implementation of our refined explanation representations and the implementation of the selected model-agnostic methods[1].

The remainder of the paper is as follows. We first provide background information and related work on local model-agnostic explanations from XAI methods and eye-tracking technology. After that, we present the task domain, dataset, and ML model used in our experiments, an overview of the four XAI methods evaluated, as well as the pursued iterative design approach. Next, we present the iterative design process results and refined explanation representations. We then describe the eye-tracking laboratory study in detail and present its results. Finally, we discuss our results and suggest possible future research directions.

2 RELATED WORK

As a result of the extensive research performed in XAI in different research communities over the last few years, researchers and practitioners have developed many innovative algorithms, visualizations, interfaces, and toolkits. For instance, some methods extract easily interpretable rules from the predictive model and present them to users as an explanation of the model's decision [21, 66, 92, 99]. Alternatively, others highlight regions of an image to indicate which pixels were influential in the model's prediction [59, 97, 101]. Several studies have surveyed the literature to provide a detailed overview of XAI by presenting the different approaches developed and providing classifications or conceptual frameworks (e.g., [2, 14, 39, 65, 68, 93]). In this section, we first introduce a classification scheme for XAI methods. We then present some of the different model-agnostic explanations proposed in the literature. Next, we discuss the use of eye-tracking in evaluating XAI explanations. Finally, we present related work that evaluates local explanations from multiple model-agnostic XAI methods.

2.1 Classification of XAI Methods

XAI methods developed by researchers and practitioners vary significantly in their approach to generate explanations. Therefore, several classification frameworks have been proposed in the literature across a set of categories. These categories used are neither mutually exclusive nor exhaustive. Instead, they represent a helpful means to compare and understand the different approaches of these XAI methods. In particular, three dimensions have commonly been used to describe explanations generated by XAI methods (1) relationship to predictive system, (2) applicable model class, and (3) explanation scope [25, 68, 87].

[1] https://github.com/miguelmezamartinez/Local-Model-agnostic-Explanations-Representations

According to their relationship to the predictive system, XAI methods can be classified into ante-hoc and post-hoc. Ante-hoc methods use intrinsically interpretable ML models for predicting and explaining [87]. These models are inherently transparent, as their parameters directly reveal how the model works (e.g., attributes' weights in linear regression models) [14, 68]. Nonetheless, the interpretability of these models is directly related to their complexity. Usually, the higher the complexity, the more difficult it is to explain its inner workings [2]. Furthermore, there is generally a trade-off between interpretability and accuracy for these models, as simpler models are usually not the most accurate ones [11]. In contrast, post-hoc methods avoid the trade-off between interpretability and accuracy by using a complex "black-box" model for predictions and a simpler model for explanations. The simpler model generates explanations after the predictive model has been trained by attempting to mimic its behavior [68, 87]. However, it is necessary to develop such post-hoc methods carefully to avoid generating easily interpretable but misleading explanations [25].

Concerning their applicability, XAI methods are classified across three degrees of portability: model-specific, model-class-specific, and model-agnostic [84]. Model-specific methods are limited to providing explanations only for the specific predictive model on which they are trained. For example, rules extracted from a decision tree model to provide explanations would not be valid for other decision tree models. Alternatively, model-class-specific methods are generalizable to provide explanations for a specific model family. Examples of model-class-specific methods can be found in computer vision, where approaches have been developed specifically to explain the behavior of neural network models [39]. In contrast, model-agnostic methods are not bound to any model or model family and can generate explanations for any model [2]. Unlike model-specific and model-class-specific methods, model-agnostic methods provide great flexibility in selecting any ML model for a given task by isolating the explanation generation from the underlying predictive model [80].

XAI methods are classified as global and local according to the scope of their explanations. Global methods provide a comprehensive, holistic model explanation, which describes the entire model behavior across all instances for a given dataset [39]. Therefore, these methods can help investigate population-level effects, such as identifying factors influencing drug consumption or climate change [2]. An example of a global method is Accumulated Local Effects (ALE) plots [5]. On the other hand, local methods provide explanations for specific instance decisions, which means that explanations are generated considering the vicinity of the instance to be explained [68]. Local explanation methods utilize the idea that even complex models expose a more simple, comprehensible behavior locally around the instance of interest [14].

2.2 Model-agnostic XAI Methods

As previously mentioned, many model-agnostic XAI methods have been proposed in the literature to provide explanations of AI systems' predictions. These methods vary significantly in their approach to generating explanations. To provide an overview of the explainable approaches used by some model-agnostic XAI methods, we rely on three categories: (1) association between antecedent and consequent, (2) contrast and differences, and (3) causal mechanisms [50, 87].

Association between antecedent and consequent includes explanation approaches that utilize item(s)-predictions relations such as influential instances [87]. These methods select particular instances of the dataset to explain the underlying data distribution and are more suitable for data humans can easily understand (e.g., images or text) [4, 68]. For example, Koh and Liang [56] utilized influence functions to identify instances on the training dataset that are more responsible for a given prediction. Furthermore, this category also includes approaches that consider the relationship between features and predictions [87]. Some methods, such as Partial Dependence Plots (DPD) [32] and Individual Conditional Expectation (ICE) plots [34], describe how features influence all model's predictions to provide global explanations. In contrast, methods such as LIME [81] analyze the influence

of features on a particular decision to provide local explanations. Alternatively, other methods, such as SHAP [62], analyze features' effects on predictions to generate local and global explanations.

The category contrasts and differences includes approaches that evaluate the similarities and dissimilarities of instances in the dataset. For instance, Kim et al. [55] proposed using representative instances called prototypes and instances not well represented by those prototypes (criticisms) to provide global explanations. Likewise, Ribeiro et al. [82] developed Anchors, a method that generates local explanations by analyzing similar instances to derive high-precision rules representing sufficient conditions for the prediction. Moreover, this category also includes approaches that utilize contrasts to present explanations. Class-contrastive counterfactual statements are a prominent example that has gained interest in the literature, as they are believed to be comprehensible, human-friendly explanations [65]. Counterfactual explanations describe a causal relationship in the form of "if X had not occurred, Y would not have occurred" [68]. Examples of model-agnostic methods that provide counterfactual explanations include AdViCE [35] and DICE [69].

Finally, the category of causal mechanisms considers approaches that generate explanations by analyzing causal relationships. Historically, researchers have used causal models to analyze the causal relationships from statistical data in an individual system or a population [44]. Some explainability approaches, such as counterfactual statements and DPD, are considered to have a causal interpretation because they analyze which changes to the input attributes lead to a given prediction [68]. Other examples of causal approaches include the work of Heskes et al. [43] and Frye et al. [33], which adapted the concept of Shapley Values to generate causal explanations.

2.3 Eye-Tracking Technology in XAI Research

Since eye-tracking was first used to investigate human visual perception over a century ago, many methods have been proposed to track eye movement and utilize it with various goals [79]. Generally, eye-tracking has been used for interactive and diagnostic purposes [28]. While users' eye movement data is used as an input modality in an interactive role, in a diagnostic role, it is used as a cue for estimating their intentions and cognitive states [29, 47]. For evaluating the usability of human-computer interfaces, eye-tracking has been identified as an objective source of data that provides an understanding of users' visual information processing [76]. In recent years, the development of more accurate and affordable eye-tracking devices and their non-intrusive data collection approach has increased their utilization in research.

To gain an overview of studies in the literature incorporating eye-tracking in the evaluation of explanations, we searched in established digital libraries for studies focusing on "explainable artificial intelligence" or "interpretable machine learning" systems, together with eye-tracking terminology. In Table 1, we present a summary of studies found along the following attributes: (1) context domain, (2) eye-tracking measures, and (3) evaluated explanations.

Bigras et al. [9] incorporated eye-tracking technology in an e-commerce context to investigate users' behavior toward recommendation agents (RA). Their research controlled for RA's transparency through model-specific feature attribution explanations. They utilized two established standard eye-tracking metrics (i.e., number of fixations and fixation duration) to investigate users' cognitive effort when interacting with RAs [61]. Likewise, Coba et al. [18] utilized eye-tracking data to analyze users' decision-making strategies when evaluating model-specific summarizations of rating distributions in an e-commerce context. Besides the number of fixations and fixation duration, they also incorporated the analysis of transitions and revisits [73] between areas of interest to investigate how users examined alternatives when making decisions.

Conati et al. [19] incorporated eye-tracking in their investigation of the value of model-specific explanations of AI-driven hints in the context of intelligent tutoring systems. They analyzed the number of fixations and fixation duration to capture how much time participants spent looking at explanations. Likewise, Polley et al. [75] incorporated eye-tracking data to evaluate their proposed model-class-specific global and local explanations

Table 1. Summary of evaluation studies using eye-tracking in XAI research (sorted by publication date).

	Context Domain				Eye-tracking Measures				Evaluated Explanations					
									Model-specif	Model-class-specific				
Studies	E-Commerce	Image Classification	Search Systems	Tutoring Systems	Fixation-based	Heatmap-based	Pupil Dilatation	Transition-based	Feature attribution	Feature attribution	Grad-CAM [86]	SIDU [70]	RISE [74]	Integrated Gradients [90]
Bigras et al. [10]	X				X				X					
Coba et al. [19]	X				X			X	X					
Conati et al. [20]				X	X				X					
Polley et al. [75]			X			X				X				
Karran et al. [52]		X					X				X			X
Muddamsetty et al. [70]		X				X					X	X	X	

in the context of search systems. They generated heatmaps from users' eye-tracking data to explore scanning patterns and investigate users' attention on regions of the provided explanations.

Karran et al. [52] utilized eye-tracking technology in the context of image classification to investigate how different model-class-specific explanation visualization strategies impact users' trust [86, 90]. Specifically, they used pupil dilatation to infer users' cognitive load when evaluating explanations. Meanwhile, Muddamsetty et al. [70] followed an alternative approach to evaluate the quality of model-class-specific visual explanations generated by an XAI method utilizing eye-tracking data [70, 74, 86]. In particular, they evaluated explanations generated by the XAI method Similar Difference and Uniqueness (SIDU), which provides visual saliency maps highlighting regions responsible for the prediction in image classification. They gathered eye-tracking data from users evaluating images for object recognition and generated heatmaps representing the users' fixations on different image regions. Subsequently, they compared these user-generated heatmaps against the visual saliency maps to evaluate the quality of the explanations

In summary, the results of our literature review indicate that research that leverages eye-tracking to evaluate explanations has mainly focused on model-specific or model-class-specific methods. However, the results indicate a lack of studies leveraging eye-tracking for evaluating explanations from model-agnostic XAI methods.

2.4 Local Model-agnostic Explanations

To gain an overview of similar work conducted in the literature so far, we searched in established digital libraries for studies that conducted evaluations with different types of participants of local model-agnostic explanations from at least two different XAI methods. Furthermore, we did not consider research that evaluated global model-agnostic explanations or evaluated model-agnostic explanations against non-model-agnostic explanations. In Table 2, we present a summary of the comparative studies found along the following attributes: (1) XAI methods, (2) evaluation context, (3) type of participants, (4) representation used, and (5) evaluation measures.

Binns et al. [10] conducted between-subjects experiments with students and users to evaluate the effect of explanation styles on perceived fairness in different contexts. They compared input influence-based explanations

Table 2. Summary of evaluation studies of local model-agnostic explanations generated by XAI methods (sorted by publication date).

	XAI Methods											Context Domain					Type of Participants				Repre-sentation			Evaluation Measures		
Studies	Anchors [82]	Case-based [26]	CluReFI [30]	Decision Boundary [41]	Demographic [6]	GAMs [32]	LIME [81]	Prototype [41]	QII [20]	Sensitivity Analysis [78]	SHAP [62]	Economy	Employment	Entertainment	Justice	Science	Data Scientists	Domain Experts	Students	Users	Adapted	Original	Textual	Behavioral (e.g., eye-tracking)	Self-reported	Performance
Binns et al. [10]		X			X		X		X	X		X	X	X					X	X			X		X	
Ribeiro et al. [82]	X						X					X			X	X			X			X			X	X
Dodge et al. [23]		X			X		X		X	X					X					X			X		X	
Kaur et al. [54]						X					X		X				X					X			X	X
El Bekri et al. [30]			X				X					X								X		X			X	
Hase and Bansal [41]	X			X			X	X					X	X					X		X		X		X	X
Jesus et al. [49]							X				X	X						X					X		X	X

(i.e., LIME and QII) [20, 81], case-based explanations [26], demographic explanations [6], and a type of counterfactual called sensitivity-based explanations [78] by using self-report measures. Their analysis shows that sensitivity-based explanations led to a significantly higher fairness perception than case-based and demographic explanations, while the difference with input influence-based explanations was not significant. Dodge et al. [23] evaluated the same explanation types with users in a criminal justice context using self-report measures and additionally manipulated the underlying classifiers' fairness. They found that sensitivity-based explanations were most effective at exposing fairness discrepancies. Nonetheless, in both studies, the authors used purely textual explanations to control the representation difference between the evaluated methods.

Ribeiro et al. [82] compared Anchors and LIME explanations in different contexts using the original explanation representations proposed in their developed libraries. In a within-subjects design using both self-report and performance measures, students were presented first predictions without explanations and then with explanations from one of the methods in a randomized order. Additionally, students had to guess the model's prediction for additional instances before and after each round of explanations. They found that students achieved a higher prediction accuracy using Anchors' explanations. For LIME, the authors found that the prediction accuracy varied drastically and, in some cases, was worse than for no explanations. Further, it took significantly less time to understand and use Anchors' explanations, which the authors attribute to their simplicity and generalizability.

Kaur et al. [54] studied data scientists' use of GAMs [32] and SHAP explanations to uncover common issues when building a model in the context of salary predictions. Their research relied on GAMs' and SHAP's original explanation representations. Their analysis using self-report and performance measures revealed that data scientists often "misuse" and over-trust interpretability tools and that the representations of explanations were hard to understand and could be misleading. Additionally, the results show that participants using GAMs had significantly higher accuracy and confidence in their understanding and lower cognitive load than those using SHAP.

El Bekri et al. [30] evaluated explanations from LIME against their proposed method CluReFI and a baseline with users in the context of bank loan applications. CluReFI extended LIME by first clustering instances and then providing LIME explanations for an instance that is the cluster's representative. Their research relied

on self-report measures to evaluate LIME's original explanation representation. Their results indicated that explanations increase trust in the system and that users preferred LIME explanations due to their balance between detail and simplicity.

Hase and Bansal [41] investigated the effect of explanations from LIME, Anchors, a type of counterfactual explanation called Decision Boundary, a Prototype model, and a composite method that combined the other four explanations[2]. Their research relied on adaptations of the original explanation representations and textual representations and utilized performance and self-report measures. These adaptations include changing LIME's color-coding for the attributes' influence, presenting Anchors' explanations as a probabilistic statement, and presenting counterfactuals as a series of attribute changes that lead to crossing the decision boundary. The authors conducted a between-subjects experiment with students in two contexts. Students had to guess the model's prediction without explanations and then with explanations of one type. They found that neither explanation type led to a significant increase in task performance.

Jesus et al. [49] evaluated explanations from LIME, SHAP, and a non-model-agnostic method on a real-world fraud detection task with domain experts (i.e., fraud analysts). They investigated whether textual explanation representations from these methods could increase domain experts' performance compared to a baseline without explanations using self-report and performance measures. Their results show that presenting no explanations resulted in the highest precision and the slowest decision time. Furthermore, LIME was the least preferred explanation type due to the low diversity of features shown in its explanations.

In summary, the results of our literature review indicate that research that evaluates local model-agnostic explanations from multiple XAI methods is scarce (i.e., seven articles in total). Studies have focused on performing these evaluations along different types of participants and contexts. Moreover, three out of seven studies solely focused on textual representations. However, none of these studies has incorporated complementary data collection methods based on non-intrusive measures, such as eye-tracking, to provide insights into how users utilize explanations. Our work aims to address this unfolding research gap by relying on eye-tracking technology.

3 DESIGN CONTEXT AND METHODOLOGY

This section presents an overview of important contextual constraints (i.e., domain, dataset, ML model, evaluated XAI methods) and the iterative design method followed in our work. First, we present the domain, dataset, and ML model used. Afterward, we describe the evaluated model-agnostic XAI methods and their out-of-the-box representations for providing local explanations. Then, we describe the iterative evaluation design, measures, and analysis strategy.

3.1 Domain, Dataset, and Model

We selected the bank loan applications domain to design and evaluate the explanation representations. Specifically, we used a scenario in which an AI system evaluates the risk of bank loan applications using a set of attributes and predicts the decision to approve or reject them. This scenario is commonly used in XAI research since bank loan decisions typically involve a notion of trust in the AI system [2, 3, 10, 15], and users are familiar with the general process of requesting a loan in a bank. Furthermore, this domain is highly relevant as financial institutions increasingly use AI systems to evaluate loan applications [10], and the resulting decisions can significantly impact loan applicants.

Moreover, we used the publicly available, open-source German Credit dataset [27], which contains 1,000 instances of loan applications, each represented by 20 attributes describing the details of the loan application and the applicant's financial and personal information. We modified the dataset to adjust it for our research goal. For instance, we decided to remove the attributes "personal status and sex" and "foreign worker" as

[2]Table 2 only includes single methods. Composite methods were not included to improve readability

Table 3. Precision, recall, F1-score, and accuracy of the neural network predictive model using SMOTE.

	Precision	Recall	F1-score
Approved	0.85	0.81	0.83
Rejected	0.60	0.67	0.63
Macro average	0.72	0.74	0.73
Weighted average	0.78	0.77	0.77
Accuracy			0.77

these are considered sensitive and are prohibited in many countries legislations as they could induce unlawful discrimination. It is worth mentioning that we decided to keep the attribute "job" to maintain the accuracy of the prediction model, even though this attribute contains some categories that provide information about the residence status of the loan applicant. We also modified the original attributes' names and descriptions to improve users' understanding. The attributes used and their descriptions are presented in Figure 23 in Appendix A.

To create the predictive model used in our work, we first performed exploratory data analysis using Jupyter notebooks and the programming language Python. In this analysis, we observed that the dataset suffers from a class imbalance, with 70% of loan applications approved and only 30% rejected. This class imbalance can result in a bias towards a majority class in the predictive model. On the one hand, tackling class imbalance in the training dataset has been shown to improve the predictive model's performance and generalizability [16]. On the other hand, it can also affect how the attributes influence the model's predictions and the explanations generated for them. Thus, we decided to evaluate three approaches to tackle this class imbalance to improve the model's performance despite the influence it can have on the explanations shown to users. The three evaluated approaches were Balance Class Weights [51] using the "compute_sample_weight" functionality of the scikit-learn library[3], as well as Random Oversampling (ROS) [98] and Synthetic Minority Oversampling Technique (SMOTE) [16] from the imbalanced-learn library[4]. For the three approaches, we utilized the default parameters provided by the libraries.

Using the popular Python deep learning libraries Keras[5] and TensorFlow[6], we performed a grid parameter search using cross-validation for a neural network model and the approaches to tackle imbalance to determine the best hyperparameters and architecture. The resulting model with the highest score was a neural network with 2-hidden layers, each with 65 and 33 neurons, and the SMOTE approach. Table 3 illustrates the predictive model's performance metrics.

Furthermore, a clustering approach was required in the evaluation of the third design iteration with users to control for the generalizability of explanations from one of the evaluated methods. Details regarding the reasoning for this implementation are presented in Section 4.3.1. We evaluated clustering algorithms and decided to implement a k-medoids approach as it led to the highest variability, thus, the most meaningful clusters concerning the predicted class [53]. The Gower distance was used to calculate the distance between the points, as it can handle numerical and categorical variables [36]. As a result, the instances of the dataset were divided into 13 cluster groups. For the evaluation of the third design iteration, similar bank loan applications were selected from the cluster with the highest average prediction score, which means that the loan applications were more likely to be rejected.

[3]https://scikit-learn.org/
[4]https://imbalanced-learn.org/stable/
[5]https://keras.io/
[6]https://www.tensorflow.org/

3.2 Selected Local Model-agnostic Methods

To select the XAI methods to be evaluated in our research, we analyzed existing methods proposed in the literature considering a set of criteria. First, we only considered methods implemented as open-source Python packages to integrate them with our predictive model. Additionally, we evaluated their relevance in the literature and among practitioners. Finally, we analyzed the type of explanation these methods provide to integrate a diverse set of approaches to our research. On this basis, the selected methods are (1) LIME[7], (2) Anchors[8], (3) SHAP[9], and (4) DICE[10]. In particular, LIME and SHAP were selected since they are popular and widely implemented in research and practice, as exemplified by toolkits such as AIX360 [7] or InterpretML [72]. Anchors was chosen because it is supposed to provide explanations that are easy for users to understand [82]. Lastly, DICE was selected due to the solid theoretical foundations for counterfactual explanations in psychological and philosophical literature. Moreover, counterfactuals are believed to be comprehensible and human-friendly [65]. In the following, we provide an overview of these methods by referring to the original studies and present the out-of-the-box representations for providing local explanations from each library.

3.2.1 LIME. LIME [81] aims to find an interpretable model locally faithful to the underlying predictive model. Artificial data points are created by drawing perturbed versions of the training data distribution to train this interpretable model. Then, a prediction function is used to compute the labels of these synthetic samples. LIME then uses these perturbed instances to train the interpretable model scaling each perturbed instance by a proximity measure so that data points closer to the instance of interest carry more weight. An example of a LIME explanation is presented in Figure 1. The explanation shows the weights of a linear model in a plot that represents each feature value's influence on the classifier's prediction.

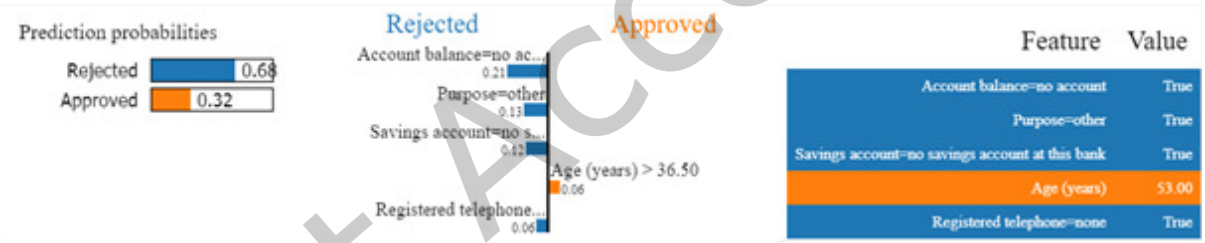

Fig. 1. Example of a LIME explanation for a credit dataset instance.

LIME's sampling strategy has received criticism due to its substantial amount of randomness, which results in a lack of robustness [100]. Additionally, it has been argued that the perturbed samples can contain many incorrectly classified instances, as the sampling process does not consider the density of the data [37], which can produce data points with high prediction uncertainty.

3.2.2 Anchors. Anchors [82] addresses some of LIME's drawbacks by explaining the logic of a predictive model with high-precision rules representing local, sufficient conditions for predictions. In other words, Anchors prescribes a set of rules for specific feature values so that altering other features does not change the prediction outcome. These rules are computed for a single instance, so they "anchor" the respective model prediction for

[7]https://github.com/marcotcr/lime
[8]https://github.com/marcotcr/anchor
[9]https://github.com/slundberg/shap
[10]https://github.com/interpretml/DiCE

this instance. Figure 2 shows an example of an Anchors' explanation. It presents the list of rules that, if fulfilled, would lead to the AI system predicting the instance as rejected 95.4% of the time.

Fig. 2. Example of an Anchors' explanation for a credit dataset instance.

The simple conditional rules Anchors generates are regarded as easy to interpret due to their clear coverage [58]. Thus, users know when to generalize an explanation given for an instance to other instances, which can help reduce the mental effort required to comprehend explanations. However, Anchors' parameters need to be carefully tuned to obtain concise rules [38]. A drawback of Anchors is that it can generate specific rules for instances close to the boundary, which can be rather complex [82].

3.2.3 SHAP. SHAP [62] provides local and global explanations by providing feature attribution scores based on the concept of Shapley values from cooperative game theory. SHAP is an adaptation from LIME that provides the Shapley values as the linear regression weights in an additive feature attribution method [62]. Figure 3 shows an example of a SHAP explanation. The base value represents the expected value of the prediction function over the training dataset, and the arrows represent the influence of each feature value toward increasing or decreasing the prediction score. These features' influences balance each other and produce the model's output score "f(x)", the classificatory prediction probability for the represented class.

Fig. 3. Example of a SHAP explanation for a credit dataset instance.

SHAP is built on a solid theoretical foundation and produces explanations that can be contrasted with explanations of other instances or a subset of instances. This contrastive property is not possible with local methods, such as LIME. Additionally, Shapley values can guarantee locally accurate and replicable explanations in theory. However, an exact computation of Shapley values is computationally expensive. Thus, approximate values are calculated, sacrificing some variability in recalculations [62].

3.2.4 DICE. In the context of AI systems, counterfactual explanations describe how the feature values of an instance would have to change to obtain a different, desirable output from the predictive model [95]. Counterfactuals cannot reveal the internal mechanisms of a model. However, they can establish some causal relationships, as the altered feature values directly influence the predictive model's outcome. Mothilal et al. [69] acknowledged that not all counterfactual explanations might be feasible for users due to the proposed modifications to the values of the features. Building on the premise that presenting a diverse set of information items to users provides benefits

in other domains of information search, the authors suggested that diversity could be beneficial when users are shown counterfactual explanations. Thus, they developed a method for generating diverse counterfactual explanations (DICE). The authors extended the work from Wachter et al. [95] as diversity and proximity of counterfactuals come with a natural trade-off. Thus, in addition to proximity, they explicitly included diversity in the counterfactuals search. DICE allows the input of relative difficulty in changing a feature by specifying feature weights. The specification of the difficulty in changing a feature can be helpful to restrict the search for counterfactuals and avoid generating explanations that include immutable feature changes. DICE presents counterfactual explanations in a tabular form, where the counterfactual changes for each feature can be visualized. Figure 4 shows an example of a DICE counterfactual explanation with only a subset of the features from the credit dataset.

	Amount (EUR)	Duration (months)	Purpose	Account Balance	Employment
Original Input Outcome: Rejected	4870	24	other	no account	between 1 and 4 years
Counterfactuals Outcome: Approved	4522	--	new car	--	--
	5080	26	--	above 200 EUR	--
	4878	23	--	--	--

Fig. 4. Example of a DICE counterfactual explanation for a credit dataset instance.

3.3 Iterative Evaluation Process

To investigate how users perceive, evaluate, and visually attend to representations of different local explanations from model-agnostic XAI methods, we relied on a design process that iteratively refined the representations and evaluated them with users. In the end, we evaluated the final version of the refined representations in a laboratory experiment using eye-tracking technology. Throughout our research journey, we adjusted the experimental design and evaluation measures according to the focus of each evaluation. This section presents an overview of this evolution and the underlying analysis strategy.

3.3.1 Evaluation Design. In general, we relied on a between-subjects experimental design to conduct three online evaluations (one in each iteration of the design process), with participants randomly assigned to one of four groups corresponding to the explanation representations for Anchors, DICE, LIME, and SHAP. These three evaluations were conducted online due to the COVID-19 pandemic. Each of these online evaluations consisted of four phases. First, participants were introduced to the bank loan application scenario, a description of the loan applications' attributes, and information on how the AI system was trained and how it provides decision recommendations to approve or reject loan applications. Additionally, they were required to answer attention-check questions to verify their understanding. Second, participants were presented with a description of the corresponding explanation representation and examples of the explanation for an approved and a rejected loan application. Third, participants performed a forward-prediction task divided into two stages, training and testing [24]. During training, participants were presented, in the same order, a set of eight loan applications (four approved and four rejected), each with the model's decision and corresponding explanation. In the testing stage, participants were asked to guess the model's prediction for eight new loan applications that displayed the application's attributes but no explanation (four approved and four rejected). The same set of loan applications was presented to participants in the same order. Fourth, participants were asked to respond to questions regarding their demographics and their evaluation of explanations in the form of short interviews or self-reported measures.

Table 4. Matrix of evaluation measures used in each evaluation round conducted with users.

Category	Measure	Iteration of Design Process (Between-subjects)			Laboratory Experiment (Within-subjects)
		1st	2nd	3rd	
Performance	Forward-prediction Score		X	X	
Self-report	Rank				X
	Satisfaction [45]			X	X
	Trust [45]		X	X	
	Understandability [64]		X	X	
	Usefulness				X
Behavioral	Fixation duration				X
	Number of fixations				X

In contrast to the experimental design for the iterative design process, for the eye-tracking laboratory experiment, we relied on a within-subjects design, showing each participant one approved and one rejected loan application with the explanation representations of each evaluated XAI method (Anchors, DICE, LIME, and SHAP). This experimental design allowed us to understand better how participants evaluated and utilized each type of explanation representation and which one they prefered. Moreover, we tracked participants' visual attention during their interaction with these explanation representations using a Tobii Eye Tracker 4C configured with a frequency of 90 Hz and its corresponding relevant research license for recording and analyzing data. In addition, we did not perform a forward-prediction task in this laboratory experiment. Thus, the laboratory experiment consisted of four phases. First, similarly to the online evaluations, participants were introduced to the bank loan scenario, a description of the loan applications' attributes, and information on how the AI system was trained and how it provides decision recommendations to approve or reject loan applications. Second, in a randomized order, participants were shown a description of one of the explanation representations followed by an approved and a rejected loan application together with that explanation representation. This process was repeated for each explanation representation. Thus, participants received a total of two explanation representations of each of the four XAI methods evaluated. Third, participants were asked to respond to questions regarding their demographics and their evaluation of explanations using self-reported measures. Fourth, semi-structured interviews were conducted with participants (see Figure 25 in Appendix A).

3.3.2 Evaluation Measures. The evaluation measures were adjusted throughout the iterative design process and the laboratory experiment according to the focus of the evaluation. Table 4 presents a summary indicating which measures were used in each user evaluation round. For more details on the evaluation measures, see also Table 8 in Appendix A.

As an objective performance measure, we utilized the number of correct guesses of the model's prediction each participant made during the forward-prediction task. This forward-prediction score has been commonly used in research as a means to investigate the quality of explanations [12, 17, 41, 77, 82]. The idea behind this measure is that participants first build a mental model of how the ML predictive model makes decisions by observing the explanations for those decisions in a training phase. Afterward, they apply this mental model to estimate the predictive model's decision on new instances in a testing phase.

We also incorporated subjective self-reported measures of constructs commonly utilized in XAI research. These self-reported measures are collected using Likert scales to capture participants' agreement with a series of statements representing each construct. In the evaluations performed in our work, we relied on the following constructs established in the literature: (1) satisfaction [45], (2) trust [45], and (3) understandability [63]. We

also utilized custom self-reported scales to measure participants' perceived usefulness of each explanation representation and their rank according to their preference.

Finally, we integrated participants' eye movement data for our laboratory experiment to derive two behavioral measures commonly used in eye-tracking research [61, 70, 75], fixation duration and the number of fixations. Similarly to Polley et al. [75], we utilized these measures to investigate which regions of the explanation representations received more attention from users. Specifically, we compared the number of fixations and fixation duration between the explanation representations and generated heatmaps to analyze users' visual attention focus. In addition, we combined these behavioral measures with self-reports and interviews to better understand how users utilize and perceive the refined representations from our iterative design process.

3.3.3 Analysis Strategy. Throughout the evaluations of the iterative design process and in the laboratory experiment, we performed a series of tests to investigate if there were any statistically significant differences across the performance, self-reported and behavioral measures. To perform these statistical analyses, we relied on SPSS[11].

In the between-subjects online evaluations performed in the iterative design process, participants evaluated one of the four explanation representations. For these evaluations, we utilized a combination of self-reported and performance measures. While the self-reported measures were collected using 7-point Likert scales, the forward-prediction score ranged from zero to eight. We analyzed these measures separately in multivariate analyses for self-reported measures and univariate analyses for the forward-prediction score. Nevertheless, the parametric one-way MANOVA and one-way ANOVA assumptions were not met. So instead, we analyzed each self-reported measure and the forward-prediction score with the nonparametric test Kruskal-Wallis provided by the "NPTESTS" function in SPSS. All post-hoc pairwise comparisons for the different explanation representations were performed automatically by SPSS using a Bonferroni correction.

In contrast, participants evaluated each explanation representation in the within-subjects laboratory experiment. We used self-reported and eye-tracking measures and self-reported control variables for this evaluation. Multiple measurements were created in related groups across the within factors for each of the self-reported and eye-tracking measures. Due to differences in measures' scales, we ran independent repeated-measures univariate analyses with each measure as a dependent variable and the related groups as the within-subjects factor. Additionally, control variables were incorporated as covariates when supported by the statistical model. We first checked for common assumptions necessary in parametric tests. These included no significant outliers, normality, and sphericity. Depending on the number of within factors, we utilized a one-way repeated measures ANCOVA or a two-way repeated measures ANOVA when the necessary assumptions were met. These tests were run using SPSS General Linear Models (GLM) function, which automatically performed all necessary post-hoc pairwise comparisons with a Bonferroni correction. We used nonparametric Friedman tests for each measure in case one or more assumptions were violated. We followed up with a pairwise Wilcoxon signed-rank test with a Holm-Bonferroni correction when statistically significant differences were found for the within-subjects factors [46].

4 RESULTS OF THE ITERATIVE DESIGN PROCESS

To increase the comparability of the out-of-the-box explanation representations from the selected methods and control for any confounding factors due to their different explanation approaches, we refined them in an iterative design process involving users. We followed the strategy proposed by Iivari [48], which aims to provide a new solution to a general problem identified by researchers. In this strategy, although researchers may be informed about some specific problems, they face uncertainties regarding the most appropriate general solution. As a result, they must identify potential users to develop and evaluate conceptual artifacts. In our work, the iterative design process, which adhered to the methodology presented in Section 3, provided comparable representations

[11] https://www.ibm.com/products/spss-statistics

after three iterations. In the following, we describe the iterative refinements and evaluations from the design process. Additionally, we provide a summary with the most relevant information.

4.1 First Design Iteration

First, we analyzed the comparability of the explanation representations presented in Section 3.2. We realized that an evaluation using these representations presents many challenges due to the differences in the amount of information presented, color coding, or layout, which could introduce additional confounding factors in the analysis. To tackle these issues, we performed a design workshop with three data scientists from an IT company and three human-computer interaction (HCI) researchers. This workshop analyzed each method's approach to generating explanations and the resulting representations. Specifically, we examined the representation style, amount of information, color coding, and terminology used. We discussed the overlaps these explanation representations had in the mentioned criteria and proposed design modifications that would allow us to increase their comparability while reducing potential confounding factors due to their differences.

4.1.1 Representation Refinement. LIME's refined representation was based on a simplified version of the matplotlib bar plot provided by its library. A dedicated bar plot area allowed us to separate the attributes' details from the bars and place them outside the plot on the y-axis to increase their readability. We maintained the class labels (i.e., approved and rejected) at the top so users could identify each attribute's influence towards a class. We also maintained each attribute's influence value next to each bar to increase their comparability. We replaced LIME's standard color coding with SHAPs highlighting attributes contributing to approval in blue and rejection in red. Figure 5 shows the first design for LIME's explanation representation.

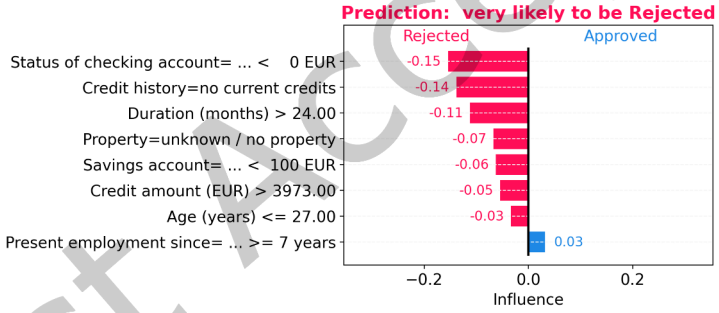

Fig. 5. First design of LIME's explanation representation.

The same bar plot was used as a baseline to refine SHAP's representation. We maintained the stacked bars representing the influence of all attributes at the top of the plot from the original representation. Like LIME's representation, to improve readability, we placed the bars and corresponding influence values inside the plot and the attributes' details on the y-axis. We maintained the base and output values from the original representation but replaced their labels with "Base Probability" and "Decision Probability" correspondingly. In contrast to LIME, where the explanation representation for "Approved" or "Rejected" classes is the same, for SHAP, the explanation representation explicitly considers only one of the two classes. For the binary classification task in our scenario, the base and output values are different for explanations of each class, as they complement each other. As we did not want to provide an explanation with a decision probability below 0.5, we presented the representation for each class according to the model's decision. However, the color coding in SHAP's original representation is bound to the increment/decrement of the model's output score, which would present a contradictory color coding for the "positive" or "negative" contribution towards approved or rejected. To address this issue, we had

two options, invert the scale of the x-axis on the plot to reverse the increment/decrement of the probability for one of the classes or invert the color-coding for the increment/decrement according to the model's decision. We considered that the direction of the increment in the x-axis was a higher constraint and decided to have different color coding for the increment/decrement for each represented class. The first design for SHAP's explanation representation is shown in Figure 6.

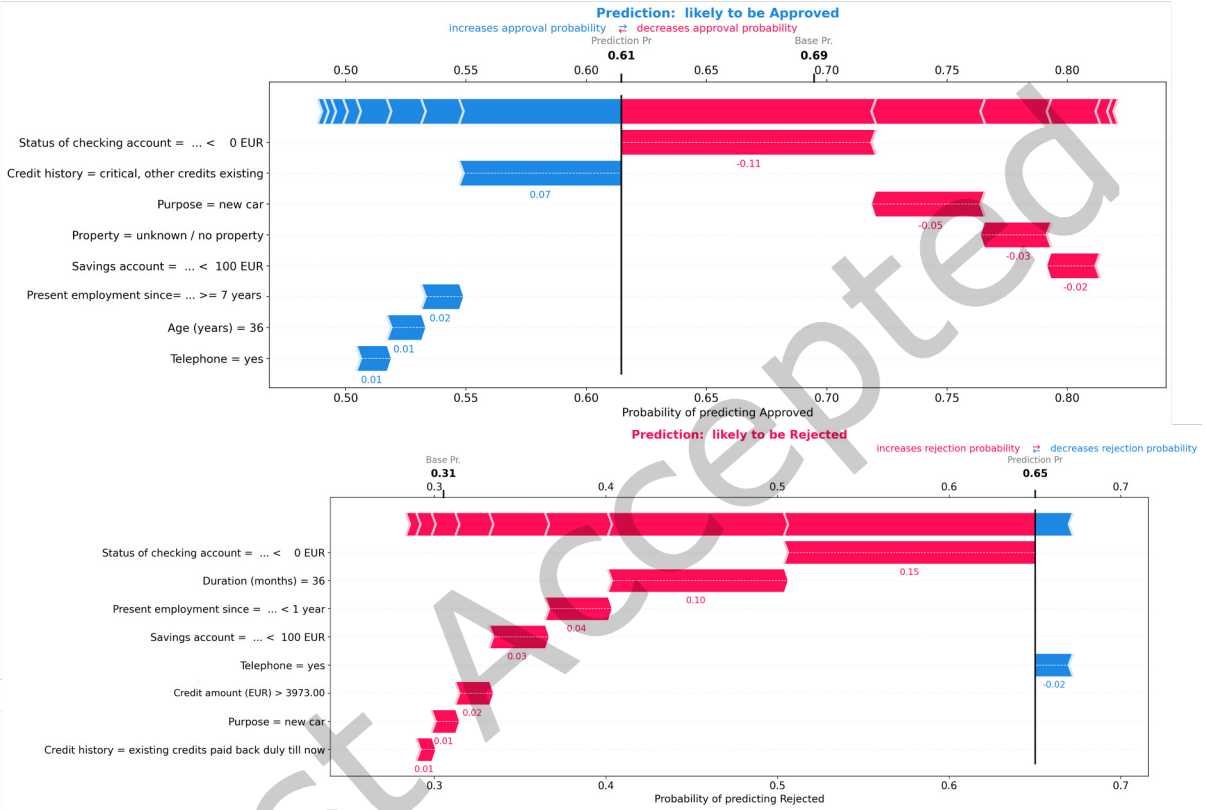

Fig. 6. First design of SHAP's explanation representations for approved and rejected loan applications.

For Anchors and DICE, we analyzed presenting rules and counterfactuals similarly to LIME and SHAP. We considered distributing the influence of attributes contained in Anchors' rules and DICE counterfactuals and showing them on a bar plot with an equal magnitude. Nevertheless, after careful analysis, we decided that this representation could lead users to interpret that each attribute's influence is independent even though the model's decision is explained as a result of the exact combination of all attributes' values shown in the rules and counterfactuals. Therefore, we used a table to present the explanation representations of both methods.

The proposed explanation representation for Anchors consisted of a table containing the rules on the left and the model's prediction on the right. We placed a header with the text "if" on top of the rules to indicate that all the conditions need to be fulfilled. Each rule was then placed on an individual row. The prediction was presented on the right part of the table in a merged cell that extended across all rules with the header "Predict". The first design for Anchors' explanation representation is shown in Figure 7.

If	Predict
Duration (months) <= 18.00 Job = skilled Registered telephone = yes	Very likely to be Approved

Fig. 7. First design of Anchors' explanation representation.

For DICE, its library lists possible counterfactuals that would lead to the alternative decision as to the model's prediction. These counterfactuals are sorted by their distance to the original instance explained by the algorithm. We decided to provide only the first counterfactual of the list, as it represents the scenario in which the model would predict the alternative class with the minimum number of attribute changes. Similar to Anchors, we presented DICE counterfactuals in a table. This table consisted of rows describing the necessary changes for each attribute in four columns: "Action", which showed "changing" for categorical attributes and "increase" or "decrease" for numerical attributes; "Attribute" with the name of the attribute; "Original Value" with the attribute's value of the loan application; and "Modified Value" with the attribute's value needed to be modified to obtain the alternative prediction. Figure 8 shows the first design for DICE's explanation representation.

The decision recommendation is "Very likely to be Approved"

The decision recommendation would be "Somewhat likely to be Rejected", if ALL the following attributes were modified:

Action	Attribute	Original Value	Modified Value
Changing	Purpose	used car	television
Changing	Housing	own	rent
Changing	Most valuable property	car	life insurance

Fig. 8. First design of DICE's explanation representation.

In addition to the individual modifications to each explanation representation, we decided to include the model's prediction probability in a text form as proposed by Cheng et al. [17]. The motivation to show the model's probability to users is to communicate the model's confidence in each decision recommendation. Thus, together with the model's decision, we presented the text "somewhat likely to be" or "very likely to be" depending on the prediction's probability.

4.1.2 Evaluation and Analysis. To evaluate the explanation representation designs proposed in the first design iteration, we invited nine university graduates to participate voluntarily in an online evaluation followed by a brief interview, as they could potentially apply for bank loans. As detailed in Section 3.3, in a between-subjects experimental design, participants evaluated one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. After the evaluation, participants were asked to compare the explanation representation they evaluated against the ones they were not shown.

Participants generally perceived the explanation representations as a useful help to understand why the AI system provided a given decision recommendation. All participants seemed to understand the basic notion of the explanation representation they evaluated and provided some feedback on how the loan application attributes were presented. Four participants recommended standardizing the use of color coding to highlight the model's decision recommendation for all explanation representations. Additionally, five participants recommended simplifying the model's decision recommendation with probability in text form for LIME and SHAP by removing the word "Prediction". Additional feedback regarding the SHAP representation was to rename the x-axis from "Probability of prediction Approved/Rejected" to "Probability of Approval/Rejection". Moreover, we received

feedback from multiple participants to make minor adjustments to the explanation representations layout for improvement. Finally, three participants brought to our attention that showing eight attributes in LIME and SHAP explanation representations could not be a fair comparison considering the number of attributes shown in Anchors' rules and DICE's counterfactuals.

4.2 Second Design Iteration

Based on the first evaluation results, we performed a second design workshop with two data scientists from an IT company and three HCI researchers. During the workshop, we analyzed the feedback provided by participants in the first design iteration and evaluated possible modifications to the explanation representations to improve their comparability further.

4.2.1 Representation Refinement. In concrete, we performed the following modifications to the explanation representations. (1) We implemented a standard color coding to the model's decision recommendation in all explanation representations. (2) Next, we simplified the probability in text form to remove the word "Prediction". (3) For Anchors, we improved the text in the table's headers to support the interpretation that all rules must be fulfilled for the model to provide the indicated decision recommendation. (4) We made minor adjustments to the axis labels and explanation representations' layout. (5) Moreover, we analyzed the number of attributes shown in each explanation representation type for all explanations in the dataset. For LIME and SHAP, it is possible to control how many attributes are shown in the explanation through configuration parameters. On the other hand, Anchors and DICE explanations introduce variability in the number of attributes shown in their explanations that depends on the instance being explained. We found that Anchors' explanations had an average of 3.34 features (SD = 3.20), while DICE's explanations had an average of 2.96 features (SD = 1.56). We decided to control the number of features shown in explanation representations to account for this. For LIME and SHAP, we limited the number of features to five. For Anchors and DICE, we selected instances from the dataset that contained between three and five features. Figure 9 shows the second design of explanation representations for Anchors, DICE, LIME, and SHAP.

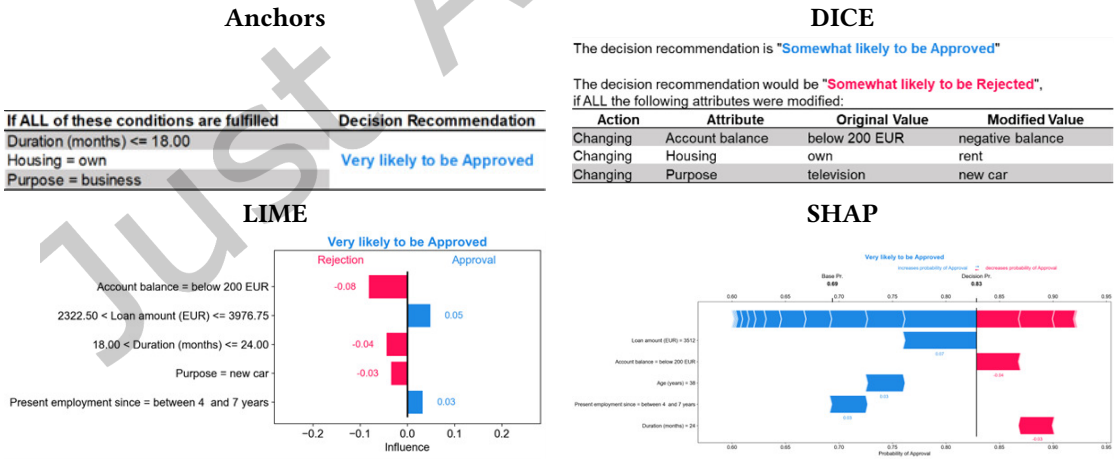

Fig. 9. Second design of explanation representations for Anchors, DICE, LIME, and SHAP.

Table 5. Descriptive statistics for the evaluation of the second iteration designs.

Variable	Group Means (SD)				
	Anchors	**DICE**	**LIME**	**SHAP**	**Total**
Trust	5.04	5.07	5.03	5.23	5.09
	(1.38)	(1.12)	(1.35)	(1.25)	(1.27)
Understandability	4.92	5.2	4.99	5.17	5.07
	(1.44)	(1.04)	(1.32)	(1.24)	(1.26)
Forward-prediction score	4.31	3.95	3.88	3.52	3.92
	(1.26)	(1.27)	(1.39)	(1.4)	(1.35)

4.2.2 Evaluation and Analysis. Next, we performed an online evaluation to evaluate the explanation representations designs from our second iteration. As detailed in Section 3.3, in a between-subjects experimental design, participants evaluated one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. After the evaluation, we asked participants to answer a questionnaire to collect their demographic information and measure their perceived understandability and trust (see Table 8 in Appendix A for details).

A total of 258 participants were recruited for the evaluation from a crowdsourcing website, as they could potentially apply for bank loans. The average time to complete the online evaluation was 28.82 minutes (SD = 10.78), and the average payment per hour was $9.72. We removed data from 23 participants as multiple instances were found where answers provided were the same word by word, which indicated that the answers could be from the same participants completing the experiment in parallel from different accounts. The final sample included 235 participants (Anchors = 58, DICE = 60, LIME = 60, SHAP = 57).

Table 5 shows the descriptive statistics of participants' perceived trust and understandability in the system and their forward-prediction score. Following the analysis strategy described in Section 3.3.3, we found that the normality assumption was violated for the multivariate analysis of trust and understandability and the univariate analysis of the forward-prediction score. Therefore, we conducted independent Kruskal-Wallis tests with trust, understandability, and forward-prediction score as dependent variables and the explanation type as the independent variable. The analyses indicate that participants perceive all explanation representations similarly, as there is no statistically significant difference between groups for trust ($\chi^2(3) = 1.751$, $p = 0.626$) or understandability ($\chi^2(3) = 0.805$, $p = 0.848$). The analysis for the forward-prediction score indicates a significant difference between groups ($\chi^2(3) = 9.963$, $p = 0.019$). Post-hoc pairwise comparisons with Bonferroni correction reveal a significantly larger score for Anchors than SHAP ($p = 0.010$). It seems that the clear coverage of the Anchors' rules allowed participants in the Anchors group to generalize the explanations observed during the forward-prediction task to achieve a higher score than participants in the SHAP group.

Through the provided open-text field, we obtained valuable feedback from participants. Several participants indicated that the explanation representations were useful, interesting, and helpful to know more about how the system makes decisions. We received multiple comments that the text showing the model's probability in the form "very likely" or "somewhat likely" was confusing, and its meaning was not clear enough. For LIME, two participants highlighted that the interpretation of the influence values shown was unclear, and they questioned whether these values should be interpreted as a percentage. For SHAP, three participants commented that the attributes with a positive and negative influence were not always on the same side of the plot. This change in position made interpreting the explanation among all eight loan applications difficult, as they would have to change back and forth the direction of the influence. For LIME and SHAP explanation representations, several participants expressed their desire to see all attributes influencing the decision instead of only the five most relevant.

4.3 Third Design Iteration

We performed a third design workshop based on the second evaluation results with two data scientists from an IT company and three HCI researchers. We observed that there seemed to be good progress in improving the comparability of the explanation representations, considering how these were perceived similarly by participants in the different groups. Nevertheless, we also identified further potential improvements from the feedback participants provided. Therefore, we performed further modifications to the explanation representations.

4.3.1 Representation Refinement. Since participants found the text showing the model's probability confusing, we replaced it with a text indicating the model's certainty in the decision recommendation as moderate or high. Considering participants' requests to see how all attributes influence the decision, we evaluated increasing the number of attributes shown for LIME and SHAP explanation representations and how it could affect the comparability with the other XAI methods. As mentioned in Section 4.2.1, controlling how many attributes are shown in the explanation through configuration parameters for LIME and SHAP is possible. With this in mind, we analyzed if it was possible to increase the amount of information on Anchors and DICE explanation representations through a new design.

For Anchors, the number of attributes depends on the rules generated by the algorithm, which can only be implicitly influenced by adapting the precision threshold of the algorithm. By changing the precision threshold, we could increase the number of features shown in the explanation. However, this could result in explanations that have specific rules with high complexity and lower coverage. An alternative would be to run the explanation algorithm multiple times to generate multiple sets of rules. Nevertheless, these different sets of rules could provide conflicting explanations. As Anchors' rules are interpreted as conditions that need to be fulfilled and would result in a particular decision, showing more than one set of rules could lead to different conditions that need to be fulfilled. On this basis, we decided not to modify the number of attributes shown in Anchors' explanations.

For DICE counterfactuals, we decided to refine the layout to present more than one counterfactual. We analyzed the distribution of the number of attributes in DICE's counterfactuals. On average, roughly seven features are presented when considering three counterfactual examples. Thus, we refined DICE's explanation representation to show three counterfactuals. For DICE's new design, we implemented a long table format in which each row represents one attribute of the bank loan application, while the columns represent the attributes' changes. Then, we highlighted the cells that contain attributes' changes using the same color coding as LIME and SHAP according to the alternative decision that the counterfactual changes would lead to. Additionally, we placed a table containing the loan application attributes' values at the left of the explanation representation as a reference. This reference would allow users to relate the attributes' changes in the counterfactuals to the current values. The attributes were grouped into loan details, financial status, and personal information for the loan application details table. For each group, a different color code and alphabetical ordering were used. The third design of DICE's explanation representation is shown in Figure 10.

The loan application attributes table was integrated for all methods at the left of the explanation representation. To adhere to the number of attributes shown in DICE's new design, we increased the number of attributes shown in the explanation representation for LIME and SHAP to seven. In contrast, the number of attributes shown in Anchors was not controlled.

Moreover, we decided to simplify Anchors' design and place the text indicating that all rules must be fulfilled at the top of the explanation representation with the rules below. Additionally, we recalculated each attribute's influence value as a percentage of the total influence from all attributes for LIME's explanations and presented them as a percentage instead. Lastly, for SHAP, we considered how changing the position of the positive and negative attributes in the plot could increase participants' mental effort. To avoid this, we decided to always present a plot with the probability of rejection, even if this probability was below 0.50. Figure 11 shows the third

AI Recommendation : Rejected
with moderate certainty

If the following attributes change : Approved
then the recommendation would be

	Applicant Information	Attribute Changes 1	Attribute Changes 2	Attribute Changes 3
Amount (EUR)	1049	1303	-	350
Duration (months)	18	-	-	-
Guarantee	none	-	guarantor	-
Purpose	used car	-	-	-
Account Balance	no account	-	-	above 200 EUR
Assets	life insurance	-	-	-
Available Income	less than 20%	-	-	-
Housing	rent	-	-	-
Loan History	paid back previous loans at this bank	-	-	-
Number of Previous Loans	1	-	-	-
Other Loans	no additional loans	-	-	-
Savings Account	no savings account at this bank	-	-	above 1000 EUR
Age (years)	21	-	-	-
Employment	less than 1 year	between 4 and 7 years	-	-
Job	skilled	-	unskilled (permanent resident)	-
Number of dependents	0 to 2	-	-	-
Residence Duration	more than 7 years	-	-	-
Telephone	none	-	yes	-

Loan Details | Financial Status | Personal Information

Fig. 10. Third design of DICE's explanation representations.

design of explanation representations for Anchors, LIME, and SHAP but omits the loan application attributes table as it has already been shown for DICE.

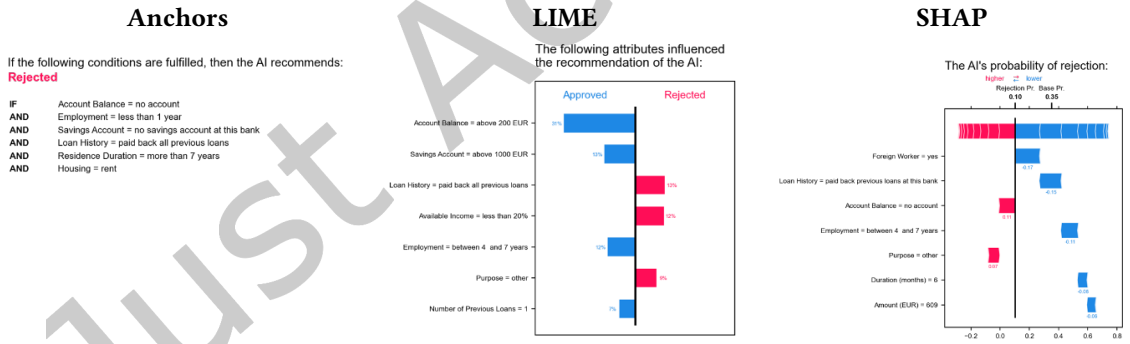

Fig. 11. Third design of explanation representations for Anchors, LIME, and SHAP.

4.3.2 Evaluation and Analysis. As detailed in Section 3.3, we conducted an online evaluation in a between-subjects experimental design, with participants evaluating one of the four explanation representations from the selected model-agnostic methods in a forward-prediction task. Moreover, we controlled for the generalizability of Anchors' explanations, which is visible in the higher forward-prediction score participants in the Anchors group achieved in the evaluation of the second iteration. To achieve this, we implemented the clustering algorithm presented in Section 3.1. This clustering approach should ensure that the local explanations seen by participants in the

Table 6. Descriptive statistics for the evaluation of the third iteration designs.

Variable	Group Means (SD)				
	Anchors	**DICE**	**LIME**	**SHAP**	**Total**
Trust	5.04	5.07	5.03	5.23	5.09
	(1.38)	(1.12)	(1.35)	(1.25)	(1.27)
Understandability	4.92	5.20	4.99	5.17	5.07
	(1.44)	(1.04)	(1.32)	(1.24)	(1.26)
Forward-prediction score	4.31	3.95	3.88	3.52	3.92
	(1.26)	(1.27)	(1.39)	(1.4)	(1.35)

training step of the forward-prediction task can help them, to a certain degree, extrapolate to the test step. Thus, we selected similar bank loan applications from the cluster with the highest average prediction score. In other words, the selected loan applications for the forward-prediction task were more likely to be rejected by the predictive model.

After the forward-prediction task, we asked participants to answer a questionnaire to collect their demographic information and measure their perceived satisfaction, trust, and understandability (see Table 8 in Appendix A for details). We recruited 261 participants from a crowdsourcing website, as they could potentially apply for bank loans. The average time to complete the online evaluation was 24.25 minutes (SD = 10.25), and the average payment per hour was $8.70. We removed data from nine participants who failed attention checks during the evaluation. The final sample included 252 participants (Anchors = 66, DICE = 63, LIME = 62, SHAP = 61).

Table 6 shows the descriptive statistics of the evaluation measures. Following the analysis strategy described in Section 3.3.3, we found that the normality assumption was violated for the multivariate analysis of satisfaction, trust, and understandability and the univariate analysis of the forward-prediction score. Therefore, we conducted independent Kruskal-Wallis tests for each measure. The analysis indicates that participants perceived all explanation representations similarly, as there was no statistically significant difference between groups for satisfaction ($\chi^2(3) = 1.415$, p = 0.702), trust ($\chi^2(3) = 1.897$, p = 0.594) or understandability ($\chi^2(3) = 1.381$, p = 0.710). In contrast to the evaluation of the second iteration, participants' trust and understandability were lower. These lower values could be explained by the selection of loan applications with a higher probability of rejection. The forward-prediction score analysis also indicates no significant difference between groups ($\chi^2(3) = 6.056$, p = 0.109), which confirmed that using our clustering algorithm to select similar loan applications helped counterbalance the generalization of Anchors' explanations.

In contrast to the evaluations in the first and second design iterations, we received no feedback regarding the design of the explanation representations from participants. On the contrary, several participants indicated that they found the explanation representations understandable and well-designed.

4.4 Summary of the Iterative Design Process

Throughout the iterative design process, we refined the out-of-the-box explanation representations from Anchors, LIME, DICE, and SHAP to increase their comparability and control for confounding factors that their original representations could induce. In the second and third online evaluations, we observed similar levels of perceived satisfaction, trust, and understandability among the groups. These results indicate that all explanation representations can, to a certain degree, help users understand how the AI system makes decisions. Nevertheless, since each participant evaluated only one of the four explanation representations in all evaluations, it was unclear which explanation type they would prefer and how they would utilize them.

Considering the results obtained in the third design iteration, we decided to terminate our design process and utilize the third explanation representations' designs shown in Figure 10 and 11 for our laboratory experiment using eye-tracking technology presented in the following section.

5 EYE-TRACKING LABORATORY STUDY

As detailed in Section 3.3, we evaluated our iteratively refined representations of local model-agnostic explanations for Anchors, DICE, LIME, and SHAP in a laboratory experiment with a within-subjects design incorporating eye-tracking technology. In contrast to the previous between-subjects online evaluations in the iterative design process, this experimental design allowed us to understand better how satisfied participants were with each type of explanation. Moreover, by analyzing users' eye-tracking data, we investigated how they visually attend to the information provided by the explanation representations, providing insights into how they utilize them. After the evaluation, we asked participants to answer a questionnaire to collect their demographic information and other control variables (i.e., domain knowledge, ML knowledge, programming knowledge, gender, and study area). Additionally, we asked participants to rank the explanation representations from most to least preferred. Finally, we conducted semi-structured interviews with participants after the laboratory experiment to discuss their perceptions of the evaluated explanation representations. During these interviews, we asked them to grade the usefulness of each explanation representation. Details for the guideline used in the semi-structured interviews can be found in Figure 25 in Appendix A.

We relied on the results of the third design iteration of the explanation representations for the evaluation in this experiment (see Figure 10 and 11). However, we faced challenges in distinguishing the visual attention on the different elements of the explanation representations due to the accuracy of the eye-tracker. Thus, we slightly modified the explanation representations' layouts by increasing the spacing between elements to differentiate participants' visual attention on them. Specifically, we modified the explanation representations to increase the separation between some of their visual elements. For Anchors, we increased the space between the rules shown. For LIME and SHAP, we reduced the number of attributes shown from seven to six and decreased the bars' height in the plots. These modifications can be observed in Figure 15.

We recruited 22 participants through a research panel at our university. Each participant was paid 12.00 € for their participation in the laboratory experiment. The average time of the actual experiment (first stage) was 29.59 (SD = 7.83) minutes, while the average time for the semi-structured interviews (second stage) was 11.71 (SD = 1.89) minutes. We removed the data from three participants due to an incomplete recording of the sessions. Thus, we analyzed data from 19 participants. All participants except one were university students. All details regarding participants' demographic information and other control variables are included in Table 9 and 10 in Appendix A. In the following, we present the results of the laboratory experiment in three sections (1) analyses of satisfaction, usefulness, and ranks, (2) analyses of participants' eye-tracking data, and (3) results of the semi-structured interviews.

5.1 Analyses of Satisfaction, Usefulness, and Rank

Table 7 shows the descriptive statistics of the self-reported measures of the eye-tracking experiment. Following the strategy described in Section 3.3.3, we conducted repeated-measures univariate analyses with each measure as the dependent variable and the related groups for each explanation type as a within-subjects factor. When the necessary assumptions were met, we conducted one-way repeated measures ANCOVA analyses and included domain knowledge, ML knowledge, programming knowledge, technical literacy, gender, and age as covariates. In case any assumptions were violated, we used nonparametric Friedman tests.

After ensuring the necessary assumptions were fulfilled, we conducted a one-way repeated measures ANCOVA analysis with satisfaction as a dependent variable. The results show that participants' satisfaction does not

Table 7. Descriptive statistics of the self-reported measures for the eye-tracking experiment.

	Group Means (SD)			
Variable	**Anchors**	**DICE**	**LIME**	**SHAP**
Satisfaction	5.06	4.45	5.03	5.64
	(1.3)	(1.57)	(0.9)	(0.87)
Usefulness	5.08	5.30	6.50	7.08
	(2.29)	(2.85)	(1.81)	(2.08)
Rank*A lower rank is better	1.84	2.47	2.74	2.95
	(0.77)	(1.26)	(1.1)	(1.08)

significantly differ across the four explanation representations (F(3, 36) = 0.382, p = 0.766), and only participants' ML knowledge has a significant influence on their satisfaction (F(3, 36) = 3.001, p = 0.043). To visualize the effect of ML knowledge on satisfaction, we conducted a follow-up analysis with satisfaction as a dependent variable, explanation representation as a within-subjects factor, and ML Knowledge as a between-subjects factor. Figure 12 shows participants' mean satisfaction for each explanation representation across the four levels of ML knowledge. Even though there are no statistically significant differences in satisfaction across the explanation representations, it is possible to observe how participants' ML knowledge level influences satisfaction. Satisfaction is similar for all explanation representations for participants with lower ML Knowledge. In contrast, there is a higher variance in satisfaction across explanation representations with higher ML knowledge.

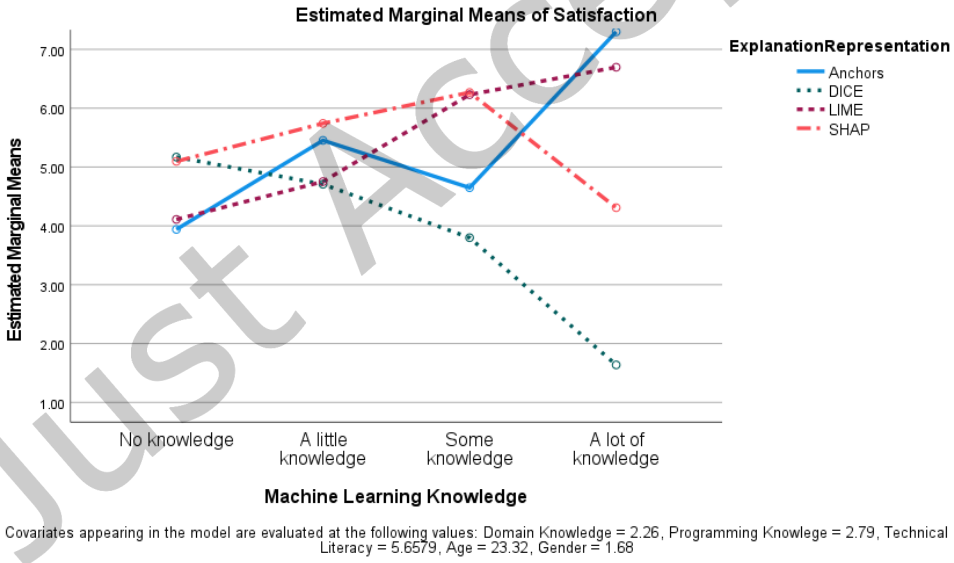

Fig. 12. Interaction effect between satisfaction and participants' ML Knowledge (error bars not included for readability).

For usefulness, the normality assumptions were violated. Thus, we conducted a nonparametric Friedman test with usefulness as the dependent variable. The results show that participants' perceived usefulness marginally differs across the four explanation representations ($\chi^2(3) = 7.190$, p = 0.066). Nevertheless, post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal no statistically significant differences between

the pairwise comparisons. As an analysis with covariates is not possible with a Friedman test, we conducted a one-way repeated measures ANCOVA with the control variables as covariates to analyze if any of them significantly affected usefulness. The results indicate no significant difference in usefulness across the explanation representations ($F(3, 36) = 1.640$, $p = 0.197$). Nevertheless, there was a significant influence of participants' gender on their perceived usefulness ($F(3, 36) = 30.188$, $p < 0.001$). To visualize the effect of gender on usefulness, we conducted a follow-up analysis with usefulness as a dependent variable, explanation representation as a within-subjects factor, and gender as a between-subjects factor. Figure 13 shows participants' usefulness for each explanation representation for females and males. It can be observed that females have higher perceived usefulness for Anchors and DICE on average. In comparison, males have higher perceived usefulness for LIME and SHAP.

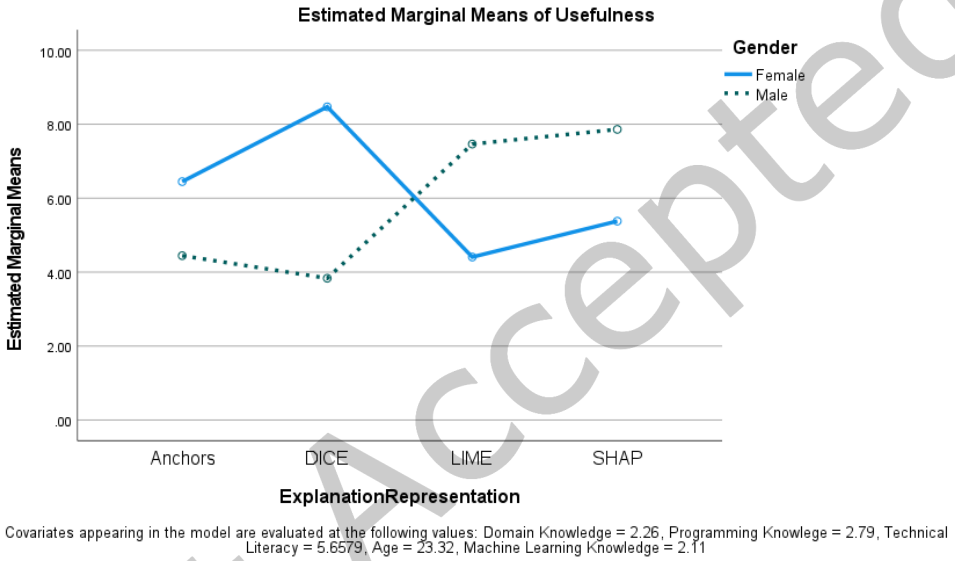

Fig. 13. Interaction effect between usefulness and participants' gender (error bars not included for readability).

A nonparametric Friedman test for participants' ranks on explanation representations shows a significant statistical difference between the related groups ($\chi^2(3) = 7.863$, $p = 0.049$). Nevertheless, post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal no statistically significant differences between the pairwise comparisons. Figure 14 shows participants' ranks for explanation representations.

5.2 Analyses of Eye-tracking Data

To analyze participants' visual attention on explanation representations from Anchors, DICE, LIME, and SHAP, we extracted the two commonly used eye-tracking measures fixation duration and number of fixations [61, 70, 75]. We utilized these measures to investigate which regions of the explanation representations received more user attention [75]. As a result, we generated eight heatmaps shown in Figure 15 by aggregating the fixations across each explanation representation and loan application decision to represent participants' visual attention focus. Users' attention levels are represented using a continuous color scale. Thus, blue stands for low attention, yellow for medium attention, and red for high attention.

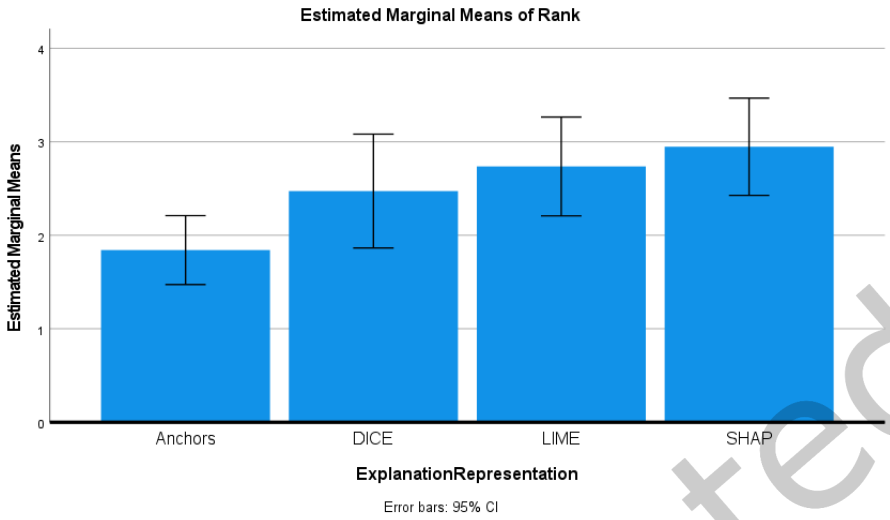

Fig. 14. Rank for explanation representations. A higher rank indicates a higher preference.

Heatmaps provide an excellent visual overview of how participants utilize the explanations during the evaluation task. An analysis can be performed to observe which regions of the explanation representations received more visual attention. Nevertheless, it is challenging to identify significant differences in visual attention between the explanation representations from observing and comparing the heatmaps visually.

Therefore, following the strategy described in Section 3.3.3, we performed statistical analyses to investigate if there were any significant differences in participants' visual attention between the explanation representations. First, we defined areas of Interest (AOI) representing regions of the explanation representations we wanted to compare. Afterward, we conducted repeated-measures univariate analyses with fixation duration or number of fixations on the AOIs as dependent variables and explanation type and loan applications as within factors. Nonetheless, considering that both measures provided similar results in almost all our analyses, we decided to report fixation duration as a primary metric and only report number of fixations when it provided additional interesting findings. Details of the description and visualization of the AOIs are found in Table 11 and Figure 24 in Appendix A.

We compared participants' visual attention on the AOIs for each explanation type in a three-level top-down approach: (1) complete visualization, including attributes' table (AOI1) and explanation representation (AOI2); (2) explanation representation (AOI2); and (3) specific elements for each explanation type. In the following subsections, we present the results of these analyses and a summary of the findings.

5.2.1 Complete Visualization. The normality assumptions were violated for fixation duration on the complete visualization (AOI1 and AOI2). Thus, we conducted a nonparametric Friedman test with fixation duration as the dependent variable. The results indicate statistically significant differences in fixation duration across the eight related groups representing explanation types and loan application decisions ($\chi^2(7) = 24.298$, $p = 0.001$). Post-hoc pairwise Wilcoxon signed-rank tests with Holm-Bonferroni correction reveal significantly lower fixation duration for Anchors approved compared to DICE approved ($p < 0.001$), SHAP rejected ($p < 0.001$), DICE rejected ($p = 0.001$), and LIME Approved ($p = 0.001$). A two-way repeated measures ANOVA analysis was conducted for fixation duration as dependent variable and explanation representation and loan decision as within factors to visualize

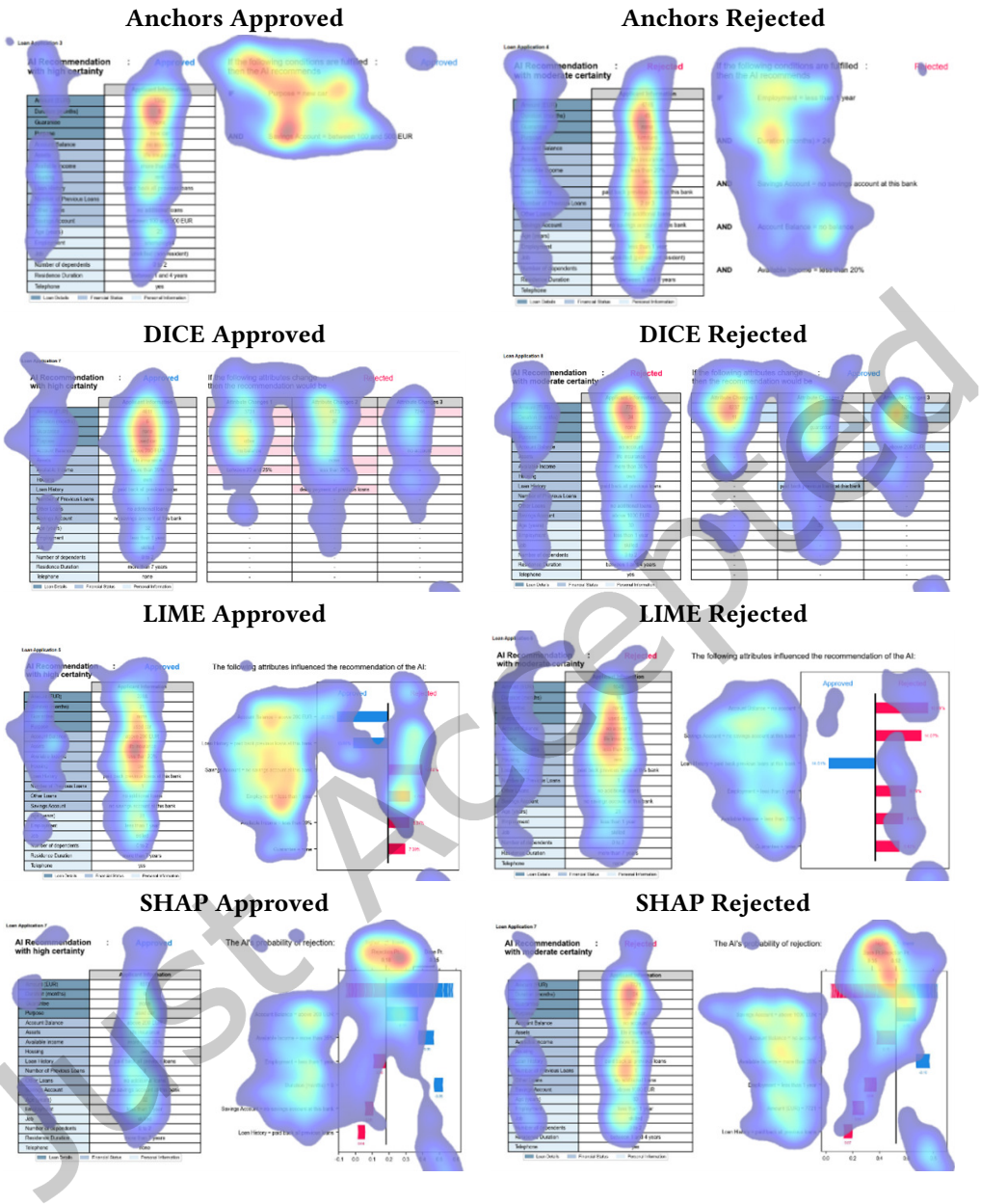

Fig. 15. Aggregated heatmaps for explanation representations by loan decision for Anchors, DICE, LIME, and SHAP.

the fixation duration in the complete visualization across the within factors. The results indicate a significant interaction effect between the within factors ($F(3,54) = 3.618$, $p = 0.019$). Figure 16 shows the interaction effect between explanation representation and loan decision on fixation duration on the complete visualization.

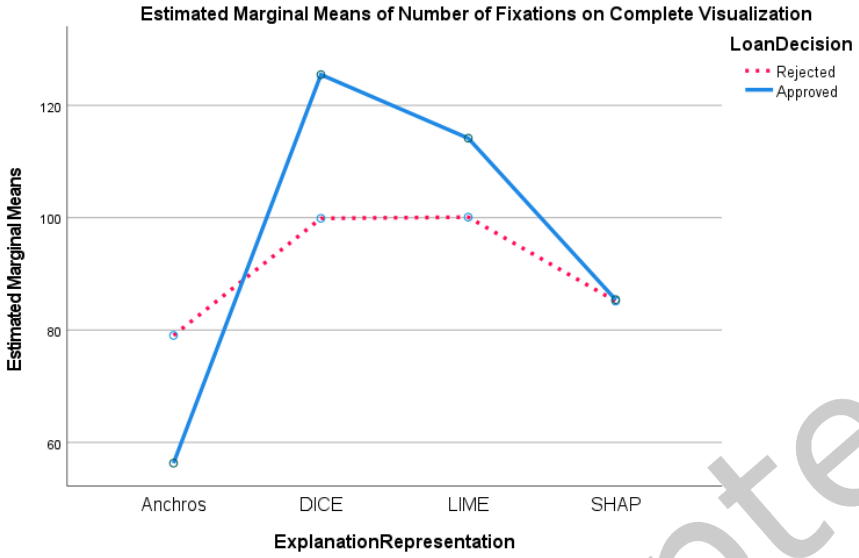

Fig. 16. Interaction effect of explanation representation and loan decision on fixation duration on complete visualization (error bars not included for readability).

These differences can be explained by the amount of information presented in each explanation representation for approved and rejected applications and the time required to analyze them. For LIME and SHAP, the number of attributes for both approved and rejected applications was six, which resulted in a similar fixation duration for both decisions. For DICE, the average number of attribute changes for rejected applications was 8.00 and approved 11.75. In contrast, for Anchors, the average number of rules in rejected applications was 4.75 and approved 2.00. These differences are also visible in Figure 15, showing fewer regions with a concentration of visual attention on the complete visualization of Anchors approved compared to the rest of the heatmaps.

5.2.2 Explanation Representation. Comparing visual attention on the explanation representation for each type presents many challenges due to the differences in information they provide. Thus, we analyzed visual attention on the explanation representations (AOI2) as a percentage of the complete visualization (AOI1 and AOI2). After verifying the necessary assumptions, we conducted two-way repeated measures ANOVA analyses with fixation duration and number of fixations as dependent variables. For fixation duration, the results show that the interaction effect between explanation representation and loan decision is not significant ($p = 0.808$). Moreover, there are statistically significant differences across explanation representations ($F(3,54) = 2.989$, $p = 0.039$) but no across loan decisions ($p = 0.535$). Post-hoc pairwise comparisons with Bonferroni correction revealed that the fixation duration for DICE is marginally lower than SHAP ($p = 0.094$). Figure 17 shows the fixation duration on explanation representation as a percentage of the complete visualization.

Similarly, the interaction effect between explanation representation and loan decision was not significant for number of fixations ($p = 0.814$). Moreover, there were statistically significant differences across explanation representations ($F(3,54) = 6.097$, $p = 0.001$) but no across loan decisions ($p = 0.259$). Post-hoc comparisons with Bonferroni correction revealed that DICE's number of fixations is significantly lower than LIME ($p = 0.024$) and SHAP ($p = 0.003$). Figure 18 shows the number of fixations as a percentage of the complete visualization for each explanation type.

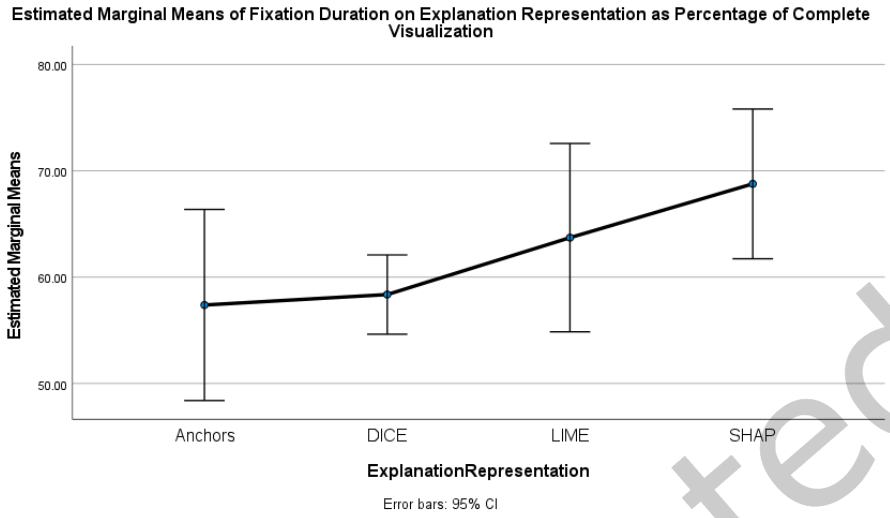

Fig. 17. Fixation duration on explanation representation as a percentage of the complete visualization.

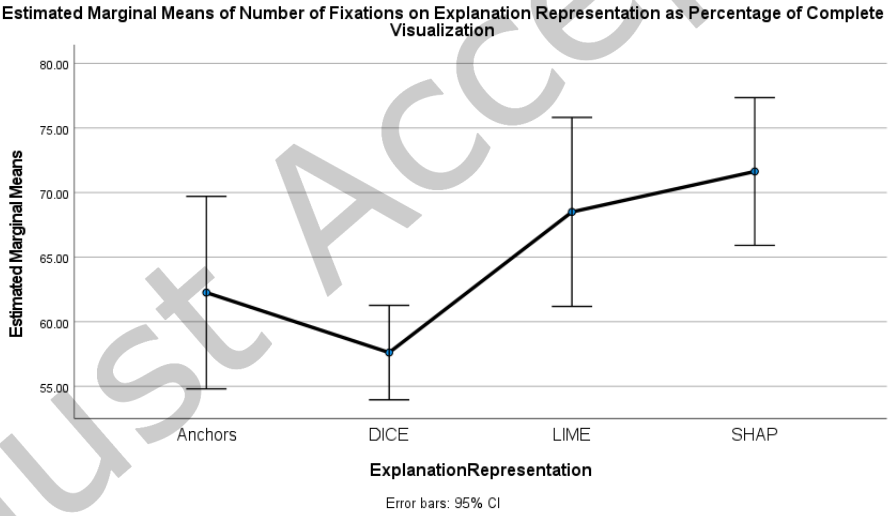

Fig. 18. Number of fixations on explanation representation as a percentage of the complete visualization.

The lower values of visual attention as a percentage of the complete visualization for DICE indicate that participants had to check the attribute's table as a reference constantly. It is also possible to observe these differences in the heatmaps of Figure 15 by contrasting the distribution of visual attention between the attributes table and the representation of DICE. This design could significantly increase users' mental effort as they must transition between the attribute's table and the representation to explore and process DICE's explanations.

5.2.3 Elements of Explanation Representations. Similarly to the approach for the explanation representations, we analyzed fixation duration on each DICE's counterfactuals as a percentage of the fixation duration in the three AOIs combined (i.e., AOI2.1, AOI2.2, and AOI2.3). After verifying the necessary assumptions, we conducted a two-way repeated measures ANOVA analysis with fixation duration as dependent variable and counterfactual and loan decision as within factors. The results show that the interaction effect between the counterfactual and the loan decision is not significant ($p = 0.616$). Moreover, there are statistically significant differences between the counterfactuals ($F(2,36) = 7.488$, $p = 0.002$). Post-hoc pairwise comparisons with Bonferroni correction revealed that fixation duration for the third counterfactual was significantly lower than the first ($p = 0.008$) and second ($p = 0.014$) (see Figure 19). When observing the heatmaps for DICE in Figure 15, it is clear that participants' visual attention focused mainly on the first and second counterfactuals. Thus, DICE's explanation representation could include only two counterfactuals to reduce information overload.

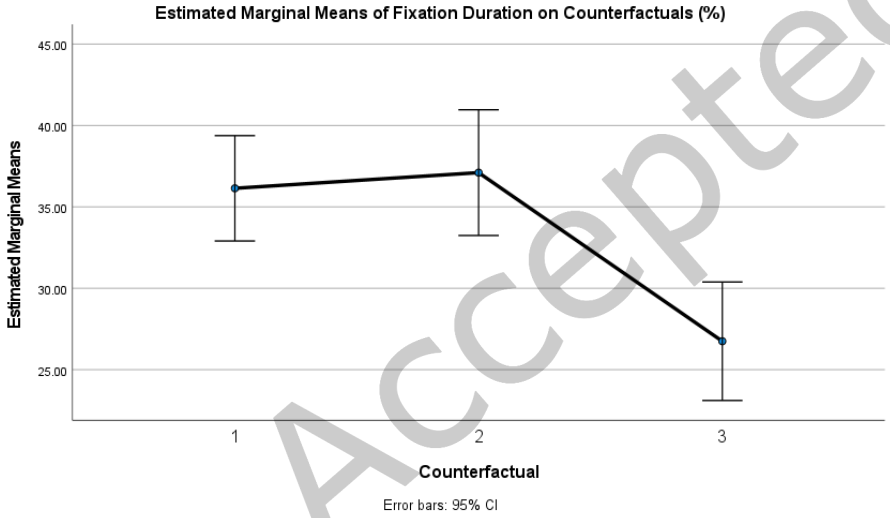

Fig. 19. Percentage of fixation duration for each counterfactual in DICE explanation representation.

Additionally, we analyzed participants' fixation duration as a percentage for LIME and SHAP by comparing top vs. bottom influencing attributes (AOI2.1 vs. AOI 2.2) and positive vs. negative influencing attributes (AOI2.3 vs. AOI2.4). After verifying the necessary assumptions, we conducted a two-way repeated measures ANOVA analysis with fixation duration as the dependent variable and explanation representation and loan decision as within factors. Regarding the comparison of top vs. bottom attributes, the results show that the interaction effect between explanation representation and loan decision is not significant ($p = 0.120$). Moreover, there is a statistically significant higher fixation duration on top attributes for LIME than SHAP ($F(1,18) = 4.581$, $p = 0.046$) but no significant difference between rejected and approved loan applications ($p = 0.887$). Figure 20 shows the fixation duration for LIME and SHAP on top attributes as a percentage. It is possible to observe these differences in visual attention in the heatmaps in Figure 15. Participants' visual attention seems more evenly distributed between the top and bottom influencing attributes for SHAP than LIME.

The results of the comparison between positive and negative attributes reveal that the interaction effect between explanation representation and loan decision is significant ($F(1,18) = 42.216$, $p < 0.001$). We conducted paired samples tests to further investigate the main effects of each within factor. For explanation representation,

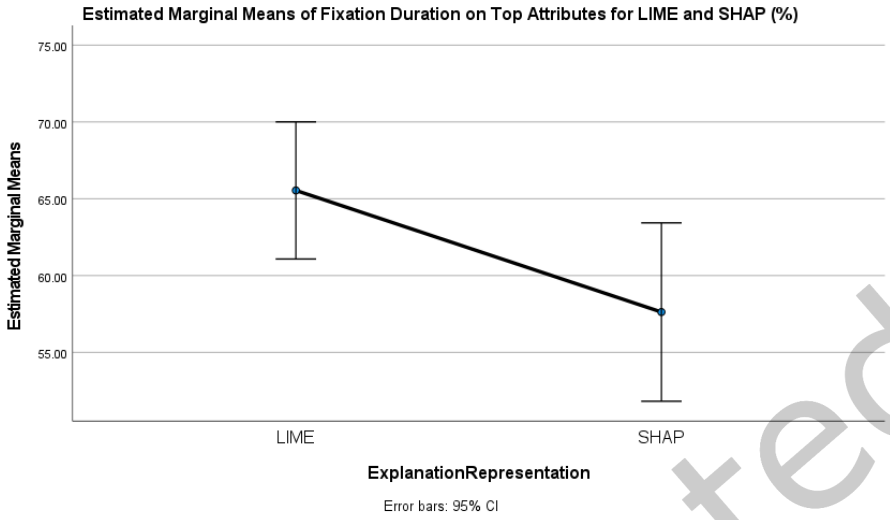

Fig. 20. Percentage of fixation duration for top influencing attributes for LIME and SHAP.

there is a significant difference between LIME rejected and SHAP rejected ($t(18) = -8.663$, $p < 0.001$) but not between LIME approved and SHAP approved ($t(18) = -1.175$, $p = 0.255$). Meanwhile, for loan decision, there are significant differences between LIME rejected and LIME approved ($t(18) = -2.672$, $p = 0.008$), as well as between SHAP rejected and SHAP approved ($t(18) = -6.543$, $p < 0.001$). Figure 21 shows the interaction effect of explanation representation and loan decision on fixation duration on the positive attributes for LIME and SHAP. This analysis reveals that the difference in positive attributes between both explanation types is more prominent for rejected loan applications.

It is possible to observe these differences in visual attention in the heatmaps in Figure 15. Participants' visual attention seems more evenly distributed between the top and bottom influencing attributes for SHAP than LIME. The heatmaps also reveal fewer regions with a concentration of participants' visual attention on positive attributes for LIME. When further analyzing the distribution of positive influencing attributes, we observed that the average for LIME rejected is 1.00, LIME approved 2.25, SHAP rejected 1.50, and SHAP approved 3.00. The dataset class imbalance explains these distributions of positive influencing attributes. In SHAP, this is reflected with a low base probability of 0.35, representing the percentage of rejected loan applications in the dataset. Therefore, there usually are more negative attributes that increase the rejection probability. Nevertheless, for LIME, this class imbalance is present only in the intercept of the local linear regression, but it is not shown as part of LIME's explanation. Consequently, for certain approved loan applications, LIME's explanations can contain a majority of negative influencing attributes, which could be counterintuitive to users (see Figure 9).

5.2.4 Summary of Findings. The eye-tracking analyses reveal lower visual attention on Anchors than on DICE, LIME, and SHAP, which can be explained by the lower number of rules presented in Anchors' explanation representations. An interpretation of this finding is that Anchors provides simpler explanations that might require lower mental effort to be processed. Nevertheless, it is unclear if this simplicity can be translated into a better understanding for users of how the AI system makes decisions. Furthermore, the analyses reveal that processing DICE's explanation representation could require a high mental effort from users due to the number of counterfactuals shown and the need to reference the attribute's table. A further refinement of DICE's explanation

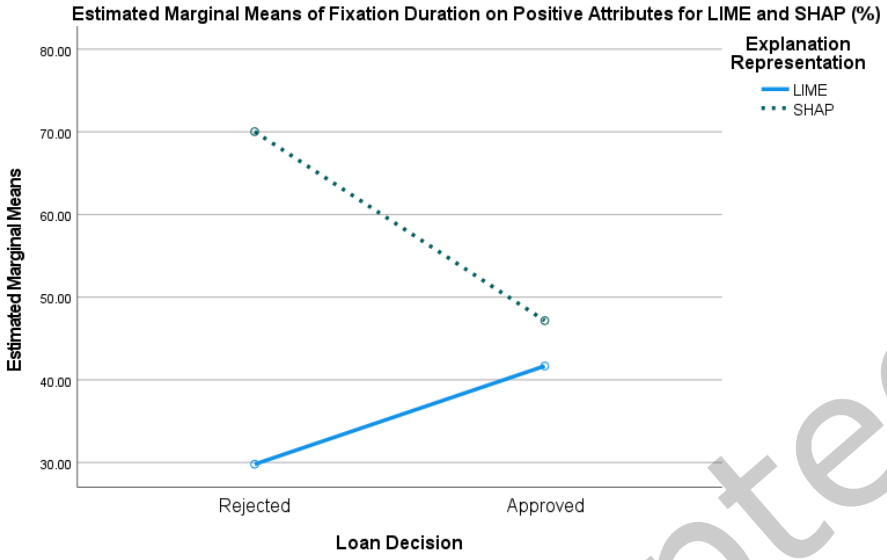

Fig. 21. Interaction effect of explanation representation and loan decision on fixation duration on positive attributes for LIME and SHAP (error bars not included for readability).

representation design could reduce the number of counterfactuals. Moreover, for SHAP, the analyses reveal a more evenly distributed participants' visual attention across the influencing attributes than LIME. Finally, the analyses reveal that LIME can generate counterintuitive explanations for imbalanced datasets by showing more attributes influencing the class contrary to the model's prediction. These counterintuitive explanations could be challenging to understand by users.

5.3 Analysis of the Semi-structured Interviews

In the following, we report our findings from the semi-structured interviews performed after the experiment regarding participants' perceptions of the AI system, attributes used, and explanation representations. Overall, our analysis shows that users process each explanation type very differently and that some factors can influence their perceptions and preferences on explanation representations. We also observed situations where participants' misinterpretation of the explanation representations negatively influenced their perception. Moreover, the interview data indicates that some explanation types might be adequate for specific situations. Figure 22 presents a summary of the most relevant findings of the interviews for each explanation representation.

5.3.1 AI System. Regarding participants' general perception of the AI system, 40% of participants indicated that the system was reliable. Additionally, 13% of participants indicated that explanations are helpful to understand how the system works and assist with decision-making. Some examples include, *"I felt that it could help a person that needs to make a decision"* (P4) and *"I know what I have to change to get my loan approved"*(P11), and *"I think it could be reliable, … there were some parameters that confused me where I personally probably would have decided differently"* (P3). Meanwhile, P10 perceived the system's reliability as limited to some explanation representations, stating: *"only [LIME] and [SHAP] are reliable recommendations"*. Four (18%) participants indicated they were unsure whether the system was reliable. P13 stated: *"I am not really sure [I can rely on it], … I need more information"*. Participants' overall relatively positive perception of the system indicates the positive effects

Anchors	DICE	LIME	SHAP
✓ Simple and easy to understand ✓ Rules are simple to follow	✓ Straightforward explanations ✓ Very simple and easy to follow ✓ Helpful to see how to change system's decision	✓ Easy to visualize and understand ✓ Correct amount of information ✓ Clearly shows how much each attribute contributed to decision • Easy to compare attributes ✓ More intuitive and simple than SHAP	✓ After explanation was understood, it was found as very helpful ✓ Stacked attributes provides a "great overview" • How all attributes interact ✓ Only explanation that provides probabilities • More confident explanations
X Misinterpretation of explanation: • System doesn't consider other attributes • Very strict and specific • Too "stupid" rules. • Possible to find "loopholes" X System was not perceived as AI	X Difference in perception when considering participants gender • More positive by females X Representation is difficult to read • Too much detail X No information on how much influence each attribute had X Just a subset of "infinite" possible counterfactuals X Unrealistic attributes' modifications	X Attributes' influence as percentage was confusing X Not clear how attributes' influence was calculated X For some approved loans, majority of attributes' influence was negative → contra intuitive • Known problem due to class imbalance	X Initially the explanation was confusing and it took some time to understand X Probabilities could be difficult to understand X Explanation is complex

Fig. 22. Summary of findings from the semi-structured interviews.

of explanations on users' perceptions. Nevertheless, factors that influence users' perceived reliability need to be considered. We further discuss this need in the following sections.

5.3.2 Attributes. 45% of participants highlighted situations where the values of the attributes seem to have a contra-intuitive influence on the system's decision. For example, for the attribute loan history, 31% of participants mentioned that the negative influence of the value *"paid back all previous loans"* on the system's decisions was unexpected. This influence contrasts with participants' expectations that having a good credit history's repayment should positively influence the system's decision, as shown by the P9 statement: *"[The system] always said that it is negative, and I never understood it"*. We analyzed the categories distribution for this attribute and observed 361 approved and 168 rejected loan applications. Thus, 68% of the loan applications with this value are approved. Nevertheless, the 168 rejected loan applications with this attribute value represent 56% of all 300 rejected applications in the dataset. As a result, the algorithm correlates the value "paid back all previous loans" for the attribute loan history with a rejection of loan applications. Researchers and practitioners need to consider the implications of providing explanations for AI systems' decisions. Explanations can reveal potential flaws or unexpected behaviors in the model, negatively influencing users' perceptions and adoption of these systems.

5.3.3 Explanation Usefulness. Moreover, 59% of participants considered that explanations' usefulness depends on the person receiving them. 27% of participants commented on the usefulness of explanations for bank employees that need to make decisions. Four (18%) participants indicated that LIME and SHAP are more helpful for decision-makers as they provide relevant information regarding each attribute's influence and relevance in relation to other attributes, as stated by P4: *"SHAP is more confident and could help people in a bank to make decisions"*. Meanwhile, four (18%) participants stated that Anchors and DICE explanations are not helpful for bank employees in some scenarios. P1 stated for DICE: *"[Bank employees] probably do not care that much about how to change the profile of the [applicant], ... this explanation is not useful for [them]"*. Similarly, P7 stated: *"In the case of [approved loan applications], showing how it can be rejected is rather useless"*. For Anchors, P9 stated: *"For someone who has to decide, [they] would not know how much everything affects the outcome"*. Furthermore, 31% of participants indicated

that certain explanations could be more useful for customers applying for bank loans. Six (27%) participants stated that DICE counterfactual explanations for rejected loan applications are very useful in this regard. P6 stated that *"you can see exactly what needs to change in order for it to be approved"*.

5.3.4 Anchors. 59% of participants stated that they perceived Anchors' explanations as simple and easy to understand. Participants stated that Anchors' *"rules were simple to follow"* (P6) and that they helped understand *"why [a loan application] was rejected or approved"* (P7). Nevertheless, two (9%) participants stated that despite the simple representation of Anchors' rules, interpreting them was difficult, including *"I felt that I was reading a flow chart or a code diagram"* (P4). P5 stated that for *"someone that does not have programming knowledge, it would be difficult to [interpret them]"*.

An interesting finding regarding Anchor's explanations is that several participants misinterpreted them, influencing their perception. Six (27%) participants perceived Anchors' explanations as *"very strict"* (P11) and *"very specific"* (P19). Their interpretation of Anchors' explanations led them to strongly criticize that the system *"only takes [those rules] into account"* (P6) and does not consider other attributes for its decision-making process. Similarly, the interpretation of two (9%) participants led them to highlight a potentially problematic situation. P19 and P21 interpreted Anchors' rules as *"too stupid because you can easily find loopholes"* (P21). They believed that for an approved loan application with only a few rules (e.g., account balance and purpose), dramatic changes to the other attributes would not influence the loan approval (e.g., unemployment). Likewise, P10 perceived that the system providing Anchors' explanations was not even AI-based because it was only following simple rules that were programmed. This analysis illustrates why participants had lower perceived usefulness and why Anchors' explanations were the least preferred.

5.3.5 DICE. For DICE explanations, we observed a clear difference in participants' perceptions when considering their gender. This difference is in line with our analysis of participants' perceived usefulness and the interaction effect with their gender in Section 5.1. As observed in this analysis, DICE's explanations had the most considerable difference in perceived usefulness for both genders. In our interviews, we found that females provided more positive comments for DICE's explanations than males and that males provided more critiques than females. Thus, to help identify these differences, we present the participants' gender in an aggregated form throughout the analysis for DICE's explanations.

Eight (36%) participants, of which four were females and four were males, indicated that DICE's explanations were *"straightforward"* (P4), *"very simple and easy to follow"* (P2). Additionally, 14 (63%) participants, of which eight were females and six were males, mentioned that DICE's counterfactual explanations are useful as they could *"see exactly what [they] need to change [the system's decision]"* (P6). They liked that DICE's counterfactuals for rejected loan applications allow them to know what the applicant *"could have done better"* (P20) to get the loan approved. On the other hand, four (18%) male participants commented that DICE's explanations provide *"too much detail"* (P1) and that *"having three columns [with counterfactuals] made it too difficult to read"* (P5). Our eye-tracking analysis supports these design critiques, where we found that the fixation duration was higher for DICE than for other explanations. Thus, DICE's explanation representation design could be improved by reducing the number of counterfactuals shown.

Moreover, three male and one female participant (18%) stated that with DICE's explanations, they *"do not know what is more important for the [system]"* (P3) as there is no information on *"how much influence did the attributes have [on the decision]"* (P18). Additionally, three male and one female participant (18%) indicated that DICE's counterfactuals are *"not enough to understand how [the system] makes decisions"* (P19) because they only show a limited number of attributes' change. P21 criticized that shown counterfactuals are *"only three examples [of changes that would lead to a different decision], but there might be hundreds or thousands of other [possible] combinations"*. P10 stated that some of the attributes' modifications proposed in the counterfactuals are not helpful as they proposed unrealistic modifications (e.g., change job or employment time). This issue with

unrealistic modifications could be addressed using DICE's library functionality to specify feature weights to restrict the search of counterfactuals and avoid generating explanations that include immutable feature changes. We decided not to provide such restrictions for our study as it could induce our own bias into the generation of counterfactuals.

5.3.6 LIME. The overall participants' perceptions of LIME explanations were positive. 40% of participants commented that they perceived LIME's explanations as *"easy to visualize and understand"* (P4). Additionally, two (9%) participants mentioned that explanations had the *"correct amount of information, not too much or too little"* (P1). Nine (40%) participants stated that LIME's explanations *"made it clear which attributes contributed to the decision"* (P12), which allowed *"an easy comparison between [the attributes]"* (P8). Despite the similarity between LIME and SHAP explanations, six (27%) participants indicated that LIME's explanations were more *"intuitive and easy to understand"* (P12) as they provided *"a better overview"* (P8).

Nevertheless, participants also highlighted some problems with LIME explanations. Four (18%) participants criticized how the attribute's influence was presented as a percentage because they interpreted it as *"not really a unit"* (P18) that *"did not mean much"* (P7) for them. They stated that LIME's explanations were missing the probability shown in SHAP as it provides a reference of how confident the decision is and how difficult it would be to change it. Moreover, five (22%) participants did not understand *"how [the percentages] were calculated"* (P6) or *"why [the attributes] affected so much or so little"* (P10). Furthermore, three (13%) participants highlighted that LIME's explanation had more attributes influencing rejection for some approved loan applications, which was very confusing. As mentioned in our analysis in Section 5.2.3, this issue is caused by a class imbalance in the dataset, having 70% of approved loan applications and 30% rejected. This imbalance is included in LIME's linear regression as the intercept, which considerably influences approving loan applications. Nevertheless, LIME's explanations do not display the intercept (see Figure 9).

5.3.7 SHAP. For SHAP, five (22%) participants indicated that they initially found the explanations confusing and that it took them some time to understand them. However, when they understood the concept, they found them very useful. Eight (36%) participants mentioned that they could clearly observe the attributes' influence and their interaction. In particular, three participants stated that having the stacked bar at the top of the explanation provided a great overview of all attributes' influence, including *"attributes are shown stacked to see how they aggregate their influence. It is almost perfect"* (P2). Additionally, seven (31%) participants stated that SHAP's explanations were the only ones that provided probabilities for the system's decision.

In contrast to the other explanations, participants consider SHAP's explanations as *"more confident"* (P4) due to their probabilistic nature. P2 stated that the base probability and the decision probabilities provided a reference to understand better the interaction of the attribute's influence in the system's decision. Nonetheless, three (13%) participants highlighted that everyone might not easily understand SHAP's explanations because they might require probability knowledge. This was the case for two participants, including P16, who stated that SHAP's explanations were *"hard to understand,"* and P4, which stated, *"what do the numbers really mean?"*.

6 DISCUSSION

In the present work, we refined and created comparable local model-agnostic explanation representations from established XAI methods following an iterative design process. Leveraging eye-tracking technology, self-reports, and interviews in the evaluation of our proposed designs helped us better understand how users process and evaluate local model-agnostic explanation representations from XAI methods. In the following, we discuss our findings and limitations and suggest possible directions for future research.

6.1 Justification

According to Swartout [91], systems must be able to explain their decisions and justify them to users in an understandable way. Swartout [91] argues that systems that fail to justify their decisions would not be accepted by their users. Our results seem to indicate that some users might require more comprehensive explanations, which provide them with a reasonable justification for the system's decision. We found that participants' preferences of explanations were influenced by their knowledge background and experience with AI systems in general and ML specifically. In particular, for participants with higher ML knowledge, there was a higher variance in satisfaction across different explanation types.

Through a qualitative analysis of data from semi-structured interviews, we further found that some participants were dissatisfied with DICE counterfactual explanations because they perceived that these explanations did not explain how the AI system had reached that decision. Counterfactual statements do not comprehensively explain the system's decisions for these participants. These counterfactual statements represent only one of the many possible scenarios that could lead to the system making an alternative decision. Additionally, for these participants, counterfactual explanations do not answer why the system had made the decision. Hereby, participants were looking for an explanation that could clarify the internal logic of the predictive model. They were interested in understanding each attribute's influence on the system decisions and how these influences interacted.

Moreover, the interaction effect of users' experience with AI systems on their preference for specific explanations indicates that researchers need to consider other potential users' characteristics when designing XAI explanations. Additional factors could influence users' need for comprehensive justification of AI system decisions.

6.2 Comparison of the selected Local Model-agnostic Methods

When refining SHAP's out-of-the-box explanation representations at the beginning of our iterative design process, we faced many challenges due to their complex design and the amount of information they provide. Additionally, we were concerned that users might not understand and trust SHAP explanation representations due to their probabilistic nature. Nevertheless, the iterative design process evaluations reveal that SHAP explanation representations were perceived similarly to Anchors, DICE, and LIME regarding trust, understandability, and satisfaction. In this line, the data analysis of the interviews revealed that despite participants' initial challenges interpreting SHAP explanations, they considered them to be very useful. Participants indicated that the base and decision probabilities provided a reference to understand better the interaction of the attributes' influences in the systems' decisions. Eye-tracking analyses supported these findings by revealing a high concentration of participants' visual attention on regions of SHAP explanation representations that show the base and decision probabilities.

According to the interviews, the rules generated by Anchors' explanations were perceived as very simple and easy to understand. These findings align with the eye-tracking analyses that revealed lower participants' visual attention on Anchors compared to the other methods. Moreover, since Anchors' rules highlight the subset of input attributes sufficient for the model to make a particular decision, they allowed participants to generalize these rules and apply them to other instances on the dataset to understand the system's decision. Nevertheless, participants in our laboratory study misinterpreted Anchors' explanations. They interpreted Anchors' explanations as a set of static rules that did not consider other attributes for the system's decision. Thus, they believed the system was not AI-based and could be fooled as loan applicants would know which attributes are "not considered" by the system to make decisions. This type of misinterpretation can strongly negatively influence users' perceptions and adoption of an AI system.

In line with findings in the literature [65], DICE's counterfactuals were found to be straightforward explanations. In the interviews, participants indicated that counterfactuals were simple and easy to follow and highlighted that seeing how to change the system's decision was very helpful. Nonetheless, they also mentioned that counterfactual explanations for approved loan applications are not very useful as they indicate modifications that would cause the loan application to be rejected. Thus, counterfactual explanations in bank loan application evaluations might be limited to explaining rejected explanations. Eye-tracking analyses revealed design flaws in DICE's explanation representations. Specifically, an analysis of participants' visual attention revealed that they need to constantly check the attribute's table as a reference to interpret the counterfactuals. Moreover, these analyses also revealed that participants' visual attention was mainly focused on the first two counterfactuals. These findings were supported by the interviews with some participants indicating that the representation was challenging to read as it had too much detail.

Overall, LIME explanations were perceived as easy to visualize and understand because they balanced the right level of detail and no information overload. Many participants preferred LIME explanations over SHAP because they found them more straightforward and intuitive. Nevertheless, eye-tracking analyses revealed that in the presence of class imbalances on the dataset, LIME could provide contra-intuitive explanations showing more attributes influencing the opposite class than the one predicted by the system (see Figure 9). In our laboratory experiment, participants were frustrated when presented with these contra-intuitive explanations. These problematic explanations could adversely affect the adoption of AI systems providing LIME explanations.

6.3 Limitations and Future Work

Our research also comes with limitations. Our iterative design process and evaluations were performed only in the specific context of bank loan applications. We selected this domain due to its relevance as financial institutions increasingly use AI systems to evaluate loan applications, and the resulting decisions can significantly impact loan applicants. Nevertheless, providing explanations of AI decisions to users is a critical issue in many other domains. Users' needs could differ for domains other than bank loan applications. Thus, future work is needed to evaluate local model-agnostic explanations in other contexts to understand these user needs and consider whether different designs are required.

Additionally, our iterative design process and evaluations focused on refining the explanation representations for a binary classification task on a tabular dataset. For example, the design of counterfactual representations requires that the attribute modifications proposed by the counterfactual statements are aligned with a table containing the attributes' names so they can be used as a reference. Nevertheless, due to the extensive research in XAI, many explainability methods have been developed to provide different types of explanations according to the type of data and ML task. Further research is needed to evaluate additional model-agnostic methods across different data types (e.g., visual or textual) and ML tasks (i.e., supervised and unsupervised).

Finally, to evaluate our explanation representation designs with users, the explanations provided during the interaction with the AI system were generated beforehand and presented a fixed number of attributes. Providing interactive explanations that allow users to explore the complete details of the explanations, such as the influence of all attributes, could enable them to understand better AI systems' decisions [64]. Therefore, future work is needed to evaluate how model-agnostic methods generate explanations and the implications of implementing them more dynamically, e.g., through increased interactivity.

7 CONCLUSION

In this work, we derive comparable local model-agnostic explanation representations from well-established XAI methods through an iterative design process. Furthermore, we use eye-tracking technology, self-reports, and interviews to understand how users process and evaluate these explanation representations. Our results indicate

that local model-agnostic explanations from different XAI methods can effectively establish satisfaction, trust, and understandability. Nevertheless, users might find some explanations more useful in specific scenarios. Moreover, our results indicate that users' preferences for model-agnostic explanations are influenced by their individual characteristics, such as gender and previous experience with AI systems. Through our work, we contribute to the ongoing research on improving the transparency of AI systems by explicitly emphasizing the user perspective on XAI.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous collaborators in the design exploration process and the reviewers for their help in improving this manuscript. The views, conclusions, and statements in this paper are those of the authors and should not be interpreted as representing any funding agency.

REFERENCES

[1] Ashraf Abdul, Christian Von Der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[3] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, and Diptikalyan Saha. 2019. Black box fairness testing of machine learning models. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, New York, NY, 625–635. https://doi.org/10.1145/3338906.3338937

[4] Aamodt Agnar and Enric Plaza. 1994. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications* 7, 1 (1994), 39–59. https://doi.org/10.3233/AIC-1994-7104

[5] Daniel W. Apley and Jingyu Zhu. 2016. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, 4 (2016), 1059–1086. https://doi.org/10.1111/rssb.12377

[6] Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied artificial intelligence* 17, 8-9 (2003), 687–714. https://doi.org/10.1080/713827254

[7] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012* (2019).

[8] Jordan Barria-Pineda, Kamil Akhuseyinoglu, Stefan Želem-Ćelap, Peter Brusilovsky, Aleksandra Klasnja Milicevic, and Mirjana Ivanovic. 2021. Explainable Recommendations in a Personalized Programming Practice System. In *International Conference on Artificial Intelligence in Education*, Vol. 12748 LNAI. Springer, 64–76. https://doi.org/10.1007/978-3-030-78292-4{_}6/TABLES/2

[9] Émilie Bigras, Pierre Majorique Léger, and Sylvain Sénécal. 2019. Recommendation Agent Adoption: How Recommendation Presentation Influences Employees' Perceptions, Behaviors, and Decision Quality. *Applied Sciences* 9, 20 (2019), 4244. https://doi.org/10.3390/APP9204244

[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–14. https://doi.org/10.1145/3173574.3173951

[11] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (10 2001), 5–32. https://doi.org/10.1023/A:1010933404324

[12] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *International Conference on Intelligent User Interfaces, Proceedings IUI*. ACM, 454–464. https://doi.org/10.1145/3377325.3377498

[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169. https://doi.org/10.1109/ICHI.2015.26

[14] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019), 832. https://doi.org/10.3390/electronics8080832

[15] Joymallya Chakraborty, Suvodeep Majumder, Zhe Yu, and Tim Menzies. 2020. Fairway: A Way to Build Fair ML Software. In *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ACM, New York, NY, 654–665. https://doi.org/10.1145/3368089.3409697

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (2002), 321–357. https://doi.org/10.1613/jair.953

[17] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–12. https://doi.org/10.1145/3290605.3300789

[18] Ludovik Coba, Markus Zanker, Laurens Rook, and Panagiotis Symeonidis. 2019. Decision-making Strategies Differ in the Presence of Collaborative Explanations: Two Conjoint Studies. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 291–302. https://doi.org/10.1145/3301275

[19] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2021. Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence* 298 (2021), 103503. https://doi.org/10.1016/J.ARTINT.2021.103503

[20] Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 598–617. https://doi.org/10.1109/SP.2016.42

[21] J. Deng and E. T. Brown. 2021. RISSAD: Rule-based Interactive Semi-Supervised Anomaly Detection. In *EuroVis 2021*. he Eurographics Association. https://doi.org/10.2312/evs.20211050

[22] Nicholas Diakopoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. 2017. Principles for Accountable Algorithms and a Social Impact Statement for Algorithms. https://www.fatml.org/resources/principles-for-accountable-algorithms

[23] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, New York, NY, 275–285. https://doi.org/10.1145/3301275.3302310

[24] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608* (2017).

[25] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018 - Proceedings*. 210–215.

[26] Dónal Doyle, Alexey Tsymbal, and Pádraig Cunningham. 2003. *A Review of Explanation and Explanation in Case-Based Reasoning*. Technical Report. Trinity College Dublin, Department of Computer Science, Dublin.

[27] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository - German Credit Data. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[28] Andrew T. Duchowski. 2002. A Breadth-first Survey of Eye-Tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (2002), 455–470. https://doi.org/10.3758/BF03195475

[29] Andrew T. Duchowski. 2017. *Eye Tracking Methodology: Theory and Practice* (third edition ed.). Springer. 1–366 pages. https://doi.org/10.1007/978-3-319-57883-5/COVER

[30] Nadia El Bekri, Jasmin Kling, and Marco F. Huber. 2019. A Study on Trust in Black Box Models and Post-hoc Explanations. In *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, Vol. 950. Springer, 35–46. https://doi.org/10.1007/978-3-030-20055-8

[31] Theodore Evans, Carl Orge Retzlaff, Christian Geißler, Michaela Kargl, Markus Plass, Heimo Müller, Tim Rasmus Kiehl, Norman Zerbe, and Andreas Holzinger. 2022. The Explainability Paradox: Challenges for XAI in Digital Pathology. *Future Generation Computer Systems* 133 (2022), 281–296. https://doi.org/10.1016/J.FUTURE.2022.03.009

[32] Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232.

[33] Christopher Frye, Chris F@faculty Ai, Colin Rowat, Ilya Feige, and Ilya@faculty Ai Faculty. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems* 33 (2020), 1229–1239.

[34] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 24, 1 (2015), 44–65. https://doi.org/10.1080/10618600.2014.907095

[35] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2021. AdViCE: Aggregated Visual Counterfactual Explanations for Machine Learning Model Validation. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 31–35. https://doi.org/10.1109/VIS49827.2021.9623271

[36] John C Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27, 4 (1971), 857–871. https://doi.org/2528823

[37] Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2019. Factual and Counterfactual Explanations for Black Box Decision Making. *IEEE Intelligent Systems* 34, 6 (2019), 14–23. https://doi.org/10.1109/MIS.2019.2957223

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *arXiv preprint arXiv:1805.10820* (2018).

[39] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *Comput. Surveys* 51, 5 (2018). https://doi.org/10.1145/3236009

[40] Kathrin Hartmann and Georg Wenzelburger. 2021. Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. *Policy Sciences* 54, 2 (2021), 269–287. https://doi.org/10.1007/s11077-020-09414-y
[41] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. ACL, 5540–5552.
[42] Mary Hayhoe and Dana Ballard. 2005. Eye Movements in Natural Behavior. *Trends in Cognitive Sciences* 9, 4 (2005), 188–194. https://doi.org/10.1016/J.TICS.2005.02.009
[43] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. 2020. Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models. *Advances in Neural Information Processing Systems* 33 (2020), 4778–4789.
[44] Christopher Hitchcock. 2018. Causal Models. In *Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
[45] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608* (2018).
[46] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70. https://www.jstor.org/stable/pdf/4615733.pdf
[47] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, Oxford.
[48] Juhani Iivari. 2015. Distinguishing and contrasting two strategies for design science research. *European Journal of Information Systems* 24, 1 (2015), 107–115. https://doi.org/10.1057/ejis.2013.35
[49] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. 2021. How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 805–815. https://doi.org/10.1145/3442188.3445941
[50] Hilary Johnson and Peter Johnson. 1993. Explanation facilities and interactive systems. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, Vol. Part F1275. ACM, 159–166. https://doi.org/10.1145/169891.169951
[51] Justin M. Johnson and Taghi M. Khoshgoftaar. 2019. Survey on Deep Learning with Class Imbalance. *Journal of Big Data* 6, 1 (2019), 1–54. https://doi.org/10.1186/S40537-019-0192-5/TABLES/18
[52] Alexander John Karran, Théophile Demazure, Antoine Hudon, Sylvain Senecal, and Pierre-Majorique Léger. 2022. Designing for Confidence: The Impact of Visualizing Artificial Intelligence Decisions. *Frontiers in Neuroscience* 16 (2022). https://doi.org/10.3389/FNINS.2022.883385
[53] Leonard Kaufman and Peter J. Rousseeuw. 2009. *Finding groups in data: an introduction to cluster analysis*. Vol. 344. John Wiley & Sons.
[54] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2019. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14. https://doi.org/10.1145/3313831.3376219
[55] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. 2016. Examples are not Enough, Learn to Criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2280–2288.
[56] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*. PMLR, 1885–1894.
[57] Kuno Kurzhals, Brian Fisher, Michael Burch, and Daniel Weiskopf. 2015. Eye tracking evaluation of visual analytics. *Information Visualization* 15, 4 (2015), 340–358. https://doi.org/10.1177/1473871615609787
[58] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, New York, USA, 79–85. https://doi.org/10.1145/3375627.3375833
[59] Will Landecker, Michael D. Thomure, Luis M.A. Bettencourt, Melanie Mitchell, Garrett T. Kenyon, and Steven P. Brumby. 2013. Interpreting Individual Classifications of Hierarchical Networks. In *Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013*. IEEE, 32–38. https://doi.org/10.1109/CIDM.2013.6597214
[60] Simon P. Liversedge and John M. Findlay. 2000. Saccadic Eye movements and Cognition. *Trends in Cognitive Sciences* 4, 1 (2000), 6–14. https://doi.org/10.1016/S1364-6613(99)01418-7
[61] Lori Lorigo, Maya Haridasan, Hrönn Brynjarsdóttir, Ling Xia, Thorsten Joachims, Geri Gay, Laura Granka, Fabio Pellacini, and Bing Pan. 2008. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology* 59, 7 (2008), 1041–1052. https://doi.org/10.1002/asi.20794
[62] Scott M Lundberg, Paul G Allen, and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 4765–4774.
[63] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th Australasian Conference on Information Systems*, Citeseer (Ed.). 6–8.
[64] Miguel Angel Meza Martínez, Mario Nadj, and Alexander Maedche. 2019. Towards an Integrative Theoretical Framework of Interactive Machine Learning Systems. In *Proceedings of the 27th European Conference on Information Systems (ECIS)*. Stockholm & Uppsala,

Sweden.
[65] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007
[66] Yao Ming, Huamin Qu, and Enrico Bertini. 2019. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 342–352. https://doi.org/10.1109/TVCG.2018.2864812
[67] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 279–288.
[68] Christoph Molnar. 2020. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* https://christophm.github.io/interpretable-ml-book/
[69] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 607–617. https://doi.org/10.1145/3351095.3372850
[70] Satya M. Muddamsetty, Mohammad N.S. Jahromi, Andreea E. Ciontos, Laura M. Fenoy, and Thomas B. Moeslund. 2022. Visual explanation of black-box model: Similarity Difference and Uniqueness (SIDU) method. *Pattern Recognition* 127 (2022), 108604. https://doi.org/10.1016/J.PATCOG.2022.108604
[71] Mohammad Naiseh, Dena Al-Thani, Nan Jiang, and Raian Ali. 2022. How Different Explanations Impact Trust Calibration: The Case of Clinical Decision Support Systems. *SSRN Electronic Journal* (2022). https://doi.org/10.2139/SSRN.4098528
[72] Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).
[73] John W. Payne. 1976. Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance* 16, 2 (1976), 366–387. https://doi.org/10.1016/0030-5073(76)90022-2
[74] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. *arXiv preprint arXiv:1806.07421* (6 2018). https://doi.org/10.48550/arxiv.1806.07421
[75] Sayantan Polley, Rashmi Raju Koparde, Akshaya Bindu Gowri, Maneendra Perera, and Andreas Nuernberger. 2021. Towards Trustworthiness in the Context of Explainable Search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2580–2584. https://doi.org/10.1145/3404835.3462799
[76] Alex Poole and Linden J. Ball. 2006. Eye Tracking in HCI and Usability Research. In *Encyclopedia of Human Computer Interaction*. IGI Global, 211–219. https://doi.org/10.4018/978-1-59140-562-7.CH034
[77] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–52. https://doi.org/10.1145/3411764.3445315
[78] Peter M Rasmussen, Tanya Schmah, Kristoffer H Madsen, Torben E Lund, Stephen C Strother, and Lars K Hansen. 2012. Visualization of nonlinear classification models in neuroimaging—Signed sensitivity maps. In *Proc. BIOSIGNALS*. Citeseer, 254–263.
[79] Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin* 124, 3 (1998), 372–422. https://doi.org/10.1037/0033-2909.124.3.372
[80] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. *arXiv preprint arXiv:1606.05386* (2016). https://doi.org/10.48550/arxiv.1606.05386
[81] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, 1135–1144. https://doi.org/10.1145/2939672.2939778
[82] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
[83] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *IUI Workshops*, Vol. 2327. 38.
[84] Marko Robnik-Šikonja and Marko Bohanec. 2018. Perturbation-Based Explanations of Prediction Models. In *Human and Machine Learning*, J. Zhou and F. Chen (Eds.). Springer, Cham, 159–175. https://doi.org/10.1007/978-3-319-90403-0{_}9
[85] Johanes Schneider and Joshua Handali. 2019. Personalized Explanation in Machine Learning: A Conceptualization. *arXiv preprint arXiv:1901.00770* (2019). https://doi.org/10.48550/arxiv.1901.00770
[86] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 618–626.
[87] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 56–67. https://doi.org/10.1145/3351095.3372870
[88] Elizabeth Stowell, Mercedes C. Lyson, Herman Saksono, Reneé C. Wurth, Holly Jimison, Misha Pavel, and Andrea G. Parker. 2018. Designing and Evaluating mHealth Interventions for Vulnerable Populations: A Systematic Review. In *Proceedings of the 2018 CHI*

Conference on Human Factors in Computing Systems. ACM, New York, NY, 1–17. https://doi.org/10.1145/3173574.3173589

[89] Benjamin Strobel, Steffani Saß, Marlit Annalena Lindner, and Olaf Köller. 2016. Do Graph Readers Prefer the Graph Type Most Suited to a Given Task? Insights from Eye Tracking. *Journal of Eye Movement Research* 9, 4 (2016), 1–15. https://doi.org/10.16910/jemr.9.4.4

[90] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 3319–3328.

[91] William R. Swartout. 1985. Explaining and Justifying Expert Consulting Programs. In *Computer-assisted medical decision making*. Vol. 2. Springer New York, 254–271. https://doi.org/10.1007/978-1-4612-5108-8{_}15

[92] Lebna V. Thomas, Jiahao Deng, and Eli T. Brown. 2021. FacetRules: Discovering and Describing Related Groups. In *2021 IEEE Workshop on Machine Learning from User Interactions (MLUI)*. IEEE, 21–26. https://doi.org/10.1109/MLUI54255.2021.00008

[93] Erico Tjoa and Cuntai Guan. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. (2019).

[94] Matt Turek. 2018. Explainable artificial intelligence (XAI). https://www.darpa.mil/program/explainable-artificial-intelligence

[95] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)* 31 (2017). https://doi.org/10.2139/ssrn.3063289

[96] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–15. https://doi.org/10.1145/3290605.3300831

[97] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2048–2057.

[98] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, and Nik Nairan Abdullah. 2014. An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013)*, Vol. 285 LNEE. Springer, 13–22. https://doi.org/10.1007/978-981-4585-18-7{_}2/COVER

[99] Jun Yuan, Oded Nov, and Enrico Bertini. 2021. An Exploration and Validation of Visual Factors in Understanding Classification Rule Sets. In *2021 IEEE Visualization Conference (VIS)*. IEEE, 6–10. https://doi.org/10.1109/VIS49827.2021.9623303

[100] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991* (2019).

[101] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 119–134.

A APPENDICES

German Credit Dataset Attributes

	Description
Amount (EUR)	Loan amount of the application in EUR
Duration (months)	Duration of loan application repayment in months
Guarantee	Information about additional people that act as a guarantee for the loan
Purpose	Purpose of the loan
Account Balance	Status of the checking account of the applicant
Assets	Most valuable available assets of the applicant
Available Income	Percentage of available income after fix costs
Housing	Information whether the housing of the applicant is owned or rented
Loan History	Information about the payment of previous loans
Number of Previous Loans	Number of previous loans of the applicant at this bank
Other Loans	Information about additional loans
Savings Account	Status of the savings account of the applicant
Age (years)	Age of the applicant in years
Employment	Duration of present employment
Job	Type of job of the applicant
Number of dependents	Number of people that depend financially on the applicant
Residence Duration	Duration living in current residence
Telephone	Information whether a telephone line is registered under the applicants name

Loan Details Financial Status Personal Information

Fig. 23. Description of German Credit dataset attributes used to train AI system.

Measures

All measured on a 7-Likert scale Strongly disagree – Strongly agree, unless stated otherwise

Table 8. Measures used in our research study.

Measure	Items
Constructs	
Trust (adapted from Hoffman et al. [45])	I am confident in the AI system. I feel that it works well.
	The recommendations of the AI system are very predictable.
	The AI system is very reliable. I can count on it to be correct all the time.
	I feel safe that when bank employees rely on the AI system, they will get the right answers.
	I am skeptical of the AI system.
	The AI system can perform the task of deciding loan applications better than a novice human user.
Understandability (adapted from Madsen and Gregor [63])	The system uses appropriate methods to provide explanations for decision recommendations.
	The system has good knowledge about this type of problem built into it.
	The system produces explanations for decision recommendations that are as good as those which a highly competent person could produce.
	The system makes use of all the knowledge and information available to it to produce explanations for decision recommendations.
Explanation Satisfaction (adapted from Hoffman et al. [45])	From the explanations, I understand how the AI system makes recommendations.
	The explanations of how the AI system makes recommendations are satisfying.
	The explanations of how the AI system makes recommendations have sufficient detail.
	The explanations of how the AI system makes recommendations seem complete.
	The explanations of how the AI system makes recommendations are useful to my goals.
Explanation Usefulness	How would you grade the usefulness of this explanation on a scale from 0 to 10?
Controls	
Domain knowledge (adapted from Cheng et al. [17])	How much experience have you had in the past with tasks similar to the evaluation of loan applications?
	1. No experience; 2. A little experience; 3. Some experience; 4. A lot of experience.
Machine learning knowledge (adapted from Cheng et al. [17])	How much knowledge of machine learning do you have?
	1. No knowledge; 2. A little knowledge – I know basic concepts in machine learning; 3. Some knowledge – I have used machine learning before; 4. A lot of knowledge – I apply machine learning frequently to my work or I create machine learning applications.
Programming knowledge (adapted from Cheng et al. [17])	How much knowledge in programming knowledge do you have?
	1. No knowledge; 2. A little knowledge - I know basic programming concepts; 3. Some knowledge - I have coded a few programs before; 4. A lot of knowledge - I code programs frequently.
Technical literacy (adapted from Cheng et al. [17])	I am confident using computers.
	I can make use of computer programming to solve a problem.
	I understand how Amazon recommends products for me to purchase.
	I use computers whenever I can.
	I understand how my credit score is calculated.
	I understand how my email provider's spam filter works.
Age	How old are you?
Gender	1. Female; 2. Male; 3 Other

Descriptive Statistics and Statistical Analyses for Eye-tracking Experiment

Table 9. Mean and standard deviation for control variables in the eye-tracking experiment.

Control Variable	Mean	SD
Domain Knowledge	2.260	0.991
Machine Learning Knowledge	2.110	0.809
Programming Knowledge	2.790	1.032
Technical Literacy	5.657	0.754
Age	23.32	2.810

Table 10. Distribution of categories for control variables in the eye-tracking experiment.

Control Variable	Category	Frequency	Percentage
Domain Knowledge	No knowledge	3	15.8
	Slightly familiar	11	57.9
	Somewhat familiar	3	15.8
	Moderately familiar	1	5.3
	Extremely familiar	1	5.3
	Total	19	100
ML Knowledge	No knowledge	4	21.1
	A little knowledge	10	52.6
	Some knowledge	4	21.1
	A lot of knowledge	1	5.3
	Total	19	100
Programming Knowledge	No knowledge	2	10.5
	A little knowledge	6	31.6
	Some knowledge	5	26.3
	A lot of knowledge	6	31.6
	Total	19	100
Gender	Female	6	31.6
	Male	13	68.4
	Total	19	100
Area of Study	Biochemistry	1	5.3
	Chemical Engineering	3	15.8
	Civil Engineering	1	5.3
	Computer Science	3	15.8
	Industrial Engineering	6	31.6
	Information Systems	1	5.3
	International Business	1	5.3
	Mathematics	1	5.3
	Mechanical Engineering	2	10.5
	Total	19	100

Definition of Eye-tracking Areas of Interest

Table 11. AOIs for each explanation representation.

	Anchors	DICE	LIME	SHAP
AOI1	Attributes Information Table			
AOI2	Explanation Representation			
AOI2.1	First rule	First Counterfactual	Top influencing attributes	
AOI2.2	Second rule	Second Counterfactual	Bottom influencing attributes	
AOI2.3	Third rule	Third Counterfactual	Negative influencing attributes	
AOI2.4	Fourth rule		Positive influencing attributes	
AOI2.5	Fifth rule			Probability area

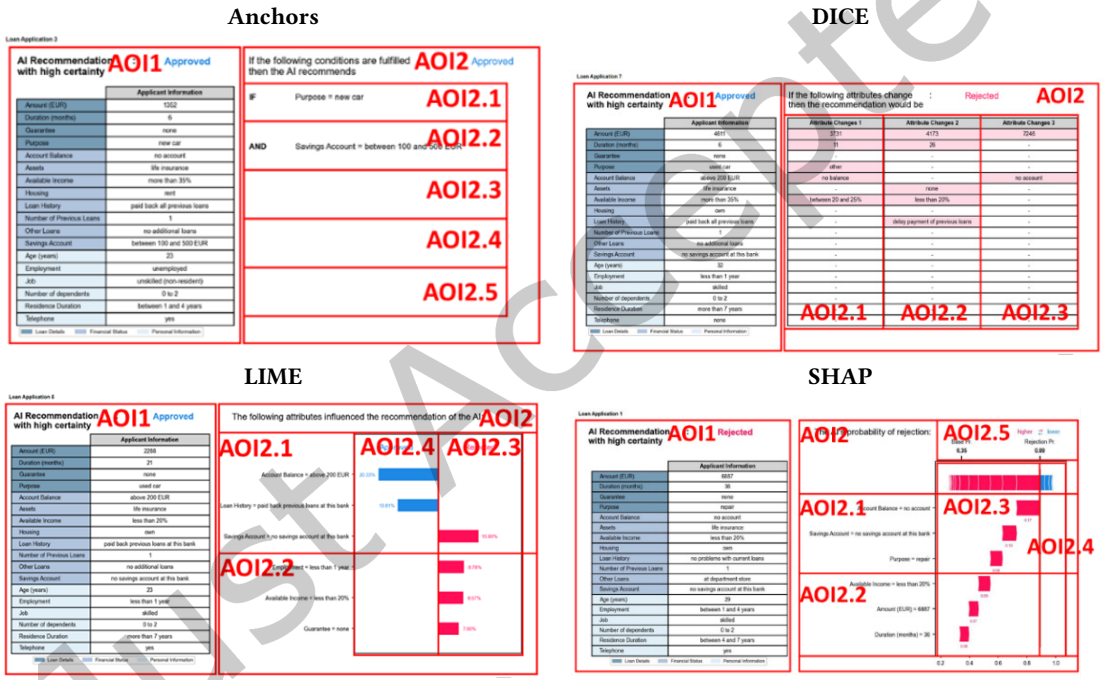

Fig. 24. Visualization of AOIs defined for the explanation representations of Anchors, DICE, LIME, and SHAP.

Guide for Semi-structured Interviews

General Experience During Experiment

- How did you feel participating in this experiment?
- Were the experiment instructions clear and easy to follow?

General Perception of the System

- How did you feel in general about your interaction with the AI system, independently of the type of explanation that the system provided?
- Do you think the system was reliable?
- How would you feel if bank employees used this system when deciding loan applications?
- Do you think the system provided fair recommendations?

General Perception on Explanations

- Do you think that providing explanations helps to understand why the system made a certain decision recommendation for a loan application?
- How would you feel if the system didn't provide any explanations for the decisions it makes?

Preference Between Explanation types

- From the four different types of explanations that the system provided to explain why it made a certain decision recommendation, how did you rank the four types of explanations according to your preference?

Preference Between Explanation types

- Why was this explanation type __ your 1st/2nd/3rd/4th preferred type of explanation?
- What did you like about this type of explanation?
- What didn't you like about this type of explanation?
- Do you think that this type of explanation was useful to understand the decisions made by the system?
- How would you grade the usefulness of this explanation on a scale from 0 to 10?

Alternative Explanations

- Could you imagine another way in which the system could provide you with an explanation of why a given decision was made?
- What information do you think could be useful and wasn't provided by any of the explanations?

Fig. 25. Guide for semi-structured interviews.