How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations

Sérgio Jesus
Feedzai, DCC-FCUP
Universidade do Porto

Catarina Belém
Feedzai
catarina.belem@feedzai.com

Vladimir Balayan
Feedzai
vladimir.balayan@feedzai.com

João Bento
Feedzai
joao.bento@feedzai.com

Pedro Saleiro
Feedzai
pedro.saleiro@feedzai.com

Pedro Bizarro
Feedzai
pedro.bizarro@feedzai.com

João Gama
LIAAD, INESCTEC
Universidade do Porto

ABSTRACT

There have been several research works proposing new Explainable AI (XAI) methods designed to generate model explanations having specific properties, or desiderata, such as fidelity, robustness, or human-interpretability. However, explanations are seldom evaluated based on their true practical impact on decision-making tasks. Without that assessment, explanations might be chosen that, in fact, hurt the overall performance of the combined system of ML model + end-users. This study aims to bridge this gap by proposing XAI Test, an application-grounded evaluation methodology tailored to isolate the impact of providing the end-user with different levels of information. We conducted an experiment following XAI Test to evaluate three popular XAI methods – LIME, SHAP, and TreeInterpreter – on a real-world fraud detection task, with real data, a deployed ML model, and fraud analysts. During the experiment, we gradually increased the information provided to the fraud analysts in three stages: *Data Only*, *i.e.*, just transaction data without access to model score nor explanations, *Data + ML Model Score*, and *Data + ML Model Score + Explanations*. Using strong statistical analysis, we show that, in general, these popular explainers have a worse impact than desired. Some of the conclusion highlights include: i) showing *Data Only* results in the highest decision accuracy and the slowest decision time among all variants tested, ii) all the explainers improve accuracy over the *Data + ML Model Score* variant but still result in lower accuracy when compared with *Data Only*; iii) LIME was the least preferred by users, probably due to its substantially lower variability of explanations from case to case.

CCS CONCEPTS

• **General and reference** → **Experimentation**; **Evaluation**; • **Computing methodologies** → *Machine learning*.

KEYWORDS

XAI, Evaluation, Explainability, LIME, SHAP, User Study

ACM Reference Format:
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 *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3442188.3445941

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
https://doi.org/10.1145/3442188.3445941

1 INTRODUCTION

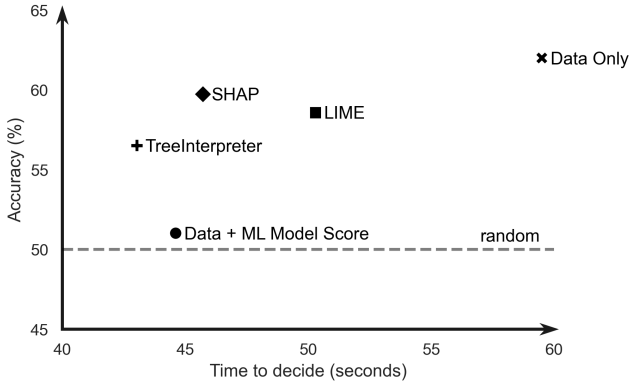


Figure 1: End-users' average decision accuracy *vs.* average time to make a decision for each variant tested in our evaluation experiment of *post-hoc* explanations. We used balanced samples of positive and negative instances, therefore, a random decision process would have 50% accuracy.

1.1 The evaluation problem in Explainable AI

The interest in ML models' explainability has been growing in the last years, as a counteractive effort to the current AI black-box paradigm, coupled with increased public scrutiny and evolving regulatory law [1–4]. However, this growth in Explainable AI (XAI)

research work has not been accompanied by effective evaluation methodologies [5]. The field is still in its early stages.

Even though every *persona* interacting with a black-box ML model may benefit of model explainability, each *persona* has a specific role, objectives, actions at disposal, background, domain knowledge, and, consequently, different explainability requirements [6–8]. As a result, the evaluation of XAI methods must be performed with the target persona and the associated task in mind [7, 8]. Notwithstanding, in seminal works of XAI methods, it is common to see introduced one or multiple *ad-hoc* evaluation setups, mostly focused on ideal explanations desiderata [9–12]. In some cases, user experiments are simulated [9] or even completely discarded from the evaluation step [13]. As a consequence, there is a lack of systematic comparison between different methods accurately and exhaustively. These reasons culminate, ultimately, in general skepticism about the reliability and usefulness of XAI methods, especially when the application is of high responsibility.

1.2 The impact of showing explanations

In this work, we focus on XAI evaluation having the end-user as target *persona*. We consider the end-user as the decision-maker, the human-in-the-loop, who usually is a domain expert, such as a judge, a doctor, or a fraud analyst. We argue that, for end-users, the value of explanations is heavily determined by how useful they are to the associated decision task and, for that reason, that their evaluation should be made by measuring their impact in the performance of the end-users. This implies involving end-users in the evaluation process, in a setup with a real task and real data. Additionally, metrics should reflect directly the users' performance, *e.g.*, how accurate the decisions are, or how fast they are made.

We propose XAI Test, an application-grounded evaluation methodology tailored to isolate the impact of gradually providing different levels of information to the end-user. A useful XAI method produces explanations that improve the overall performance of the combined system of ML model + end-user. To perform a reliable assessment, XAI Test requires testing different combinations of data, model score, and XAI methods in a real task with real end-users. Specific performance metrics must be defined (*e.g.*, accuracy or decision time), the agreement between end-users is considered on each variant, and user perception captured through questionnaires. Lastly, statistical tests are employed to detect significant differences between each variant.

1.3 The experiment

Using XAI Test, we conducted an empirical evaluation in the task of fraud detection in financial transactions. We employed three different *post-hoc* explainers and observed their impact on human-in-the-loop performance, measuring accuracy, recall, false positive rate (FPR), and decision time. We additionally collected the users' perception of usefulness, variety, and relevance of each presented explanation.

We quantified and isolated the impact of the different interacting parties in a Human-AI collaborative setting by following a three-stage evaluation approach with increased information. Figure 1 shows how the average accuracy of the decision varies with the decision time for each of the evaluated variants. We observe a clear trade-off between effectiveness and efficiency as the end-user gets access to additional ML model information. In particular, we observe that when no model-related information is shown to the end-user (*i.e., Data Only*), although slower, leads to more accurate decisions. Conversely, the accuracy obtained in the mid-level information stage (*Data + ML Model Score*) yields faster decisions but much worse accuracy - a result that is partially improved by adding model explanations.

2 RELATED WORK

In this section, we provide an overview of the current evaluation paradigm in XAI research. In particular, we briefly discuss the often considered desiderata, as well as the different techniques used to measure them. We end by enumerating a few representative state-of-the-art evaluation approaches and by describing how these fail to convey a robust analysis of the real impact of XAI methods in real-life Human-AI decision-making systems.

2.1 Desiderata

Most research work on XAI measures some kind of *proxy* of intuitive desiderata for the ideal explanation, such as **fidelity** or faithfulness [9, 14], which states that surrogate models that are used to obtain *post-hoc* explanations should be able to mimic the behavior of the explained ML model; **robustness** or stability [15], which measures whether similar input instances get similar explanations; **human-interpretability** or comprehensibility [16], which measures how easily a human interprets the result from the explanation method.

Despite being common sense that a good explanation must have high fidelity, be robust, and be intelligible, those characteristics by themselves do not say much about the actual benefit of having an explanation in a specific real-world application, nor do the measurements completely represent those characteristics.

Previous work often assumes that a model is **interpretable** because it belongs to a certain family of models – such as sparse linear models, decision trees, and rules lists [17–20], or additive models [21–23] – and the only focus when generating explanations is on the accuracy of those models. These explanations are directly derived from interpreting the ML model parameters. Most of the times, these over-simplified definitions of model intelligibility are detached from the requirements of real-world applications [24]. In general, these simpler models have much lower predictive accuracy than other more complex models, such as deep neural networks or tree ensembles. Only in a few high-stakes tasks (such as credit scoring [25]) is the complexity of an ML model viewed as an actual limitation, and only in these particular cases, there is no alternative to simpler, more intelligible models.

Several works assess **fidelity** as a measure of the quality of an explanation. Fidelity has been assessed both directly [9, 14, 26, 27], by measuring differences in predictions of the surrogate and explained models, as well as indirectly [11], by measuring how well a human can predict the output of a ML system with and without being exposed to explanations. Again, this is another metric detached from real-world impact of showing an explanation to a given persona, as it focuses on how well an XAI model approximates the function learned by the original ML model.

Other works defend the importance of **robustness**. It is measured by directly computing how much the output of an explanation method changes with its input [28, 29] or by showing the sensitivity of explanations to adversarial attacks [30]. However, these metrics are not directly related to how an explanation might help the end-user to better perform their task.

Interpretability is also assessed by measuring how approximate a XAI method explanation is to an explanation produced by a human expert [3, 31]. Those approaches are somehow restricted to tasks where the behavior of humans is intuitive, and generally close to the ground truth (such as problems in natural language processing and computer vision), but may not be suitable to complex predictive tasks based on tabular data, where the analysis has to take into account multiple features and interactions, making the task harder and less intuitive.

The way XAI desiderata is being interpreted and measured is disperse and lacking in consensus, as shown by the different methods to measure the same property. Several problems are pointed to current practices, such as non-overlapping and discordant motivations and objectives for interpretability [24], attributing the same level of interpretability to ML models originating from the same model class [32], or the lack of evaluation of XAI methods with the intended end-users [33].

Frameworks have been developed [7, 34] as an attempt on tackling the challenges of XAI evaluation, however, these frameworks are still recent and have yet to see wide adoption. The field is missing a systematic and objective way of comparing explanation methods [35, 36], which promotes research practices where each work uses customised metrics and desiderata that are thought to be the most adequate, encumbering the choice of XAI methods for a given task. This is especially important in scenarios of real-world Human-AI decision-making systems, where XAI methods may have a greater impact.

2.2 Evaluation Practices

While many *ad-hoc* evaluation setups have been used to empirically validate research on XAI methods, these either found on idyllic desiderata or overlook the human-in-the-loop and their explainability needs. In an attempt to standardize the existing XAI evaluation approaches, Doshi-Velez and Kim [5] propose a taxonomy to categorize the different types of XAI evaluation practices. In their work, the authors subdivide the evaluation practices into three distinct groups, depending on whether it resorts to humans or not and on the task they are being employed on. The first group encompasses automated evaluation on proxy tasks and is designated as *functionality-grounded* evaluation. Experiments in this category may try to simulate human behavior [11], and apply these simulations to real tasks, such as fraud detection [37]. Other works do not consider the human factor as part of the evaluation [9, 14].

Both other groups of evaluation methods use humans in the process of evaluation but differ on the task being done. If the evaluation task is a simplified proxy of a real task, the method is designated *human-grounded* evaluation, while if the task is in a real-world setting, the method is deemed *application-grounded* evaluation. These methods introduce the human component in the evaluation loop to collect feedback in the form of questionnaires, surveys, interviews, performance at the task, among others. Their focus, however, shifts from how humans perceive and interact with the explanations in *human-grounded* evaluation, to how it affects the whole system performance in *application-grounded* evaluation.

The evaluation of explanations through experimentation has been done in several past works. Most experimental studies use proxy tasks with real human subjects, *i.e.*, human-grounded experiments, such as trivia answer [38], clinical prescription simulation [16, 20], detection of deceptive reviews [39], comparison between human feedback and explainer output [3], or human prediction of model output on unseen instances based on the explanation of the model behavior [11].

By analysing the experiments conducted in other works, there is a clear gap in evaluation using real tasks with real end-users. More often than not, explanations are employed in mocked tasks, and the results obtained can not be generalized to high responsibility real-world tasks. Simulating human behavior is prone to human bias, since in many cases it depends on the developers' own intuition of the problem, and may diverge from reality, producing unrealistic results. Additionally, seldom do these experiments compare explanation methods, but rather test different visualizations or output types for these methods, which emphasizes more on the presentation rather than explanations' content.

3 EVALUATION METHODOLOGY

The evaluation of the true impact of a given explanation in the end-user experience is not an easy task. Ideally, it should be focused on objectively measuring its utility (or usefulness) in the users' decision making process. This should rely on the collection of metrics from real users while performing real tasks on real data.

We propose **XAI Test**, an application-grounded evaluation methodology that relies on realistic settings and statistical tests to robustly assess and compare the explanations' utility of different XAI methods, using metrics that correspond to the performance of the user. Rather than evaluating explainability through idyllic desiderata, we opt for evaluating it through metrics that quantify the true impact in the human decision-making.

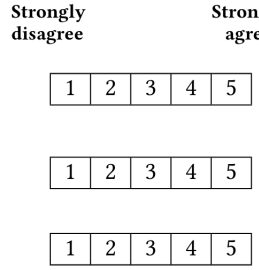
The methodology consists of the following steps: (1) formulate the hypotheses; (2) outline the experimental setup; (3) define the statistical tests to report the results with; (4) conduct the three stages of the experiment; and (5) apply statistical tests to obtained measurements.

With this methodology, we aim to find answers to a set hypotheses (*e.g., is method A more efficient than method B? Is it more accurate?*). In the case of an XAI experiment, these hypotheses are related to the utility of the explanations and how they impact the end result of a given task. To support or reject the formulated hypotheses, it is necessary to objectively measure users' performance at the task (*e.g.*, through accuracy, or decision time). It is also important do define other elements of the experiment, including the explainers, ML models, corresponding configurations to test, number of users that partake on the experiment, datasets, and other task-specific details, such as experiment scheduling and used software. Equally important for ensuring a robust evaluation is the confidence of the reported results. To this end, we define the appropriate statistical tests as well as their parameters, which are

significance level, statistical power, effect size, and sample size. A prior knowledge of the distributions is required to choose these parameters. In Section 3.3, we elaborate on the choices made in terms of hypothesis testing. The ensuing step is then to conduct the experiments in a way that isolates the impact of explanations in the decision making process. For this reason, we advocate for the execution of, at least, three stages, each providing added levels of information: (1) *Data only*, (2) *Data + ML Model Score*, and (3) *Data + ML Model Score + Explanations*. Finally, the last step of the proposed methodology concerns the collected results and their analysis.

The following sections describe the methodology employed in the evaluation of XAI methods. This includes the way explanations are employed, the measured metrics, and the battery of statistical tests to determine any significant difference.

Figure 2: Questionnaire performed to the users after each instance with explanation.



3.1 Metrics Choice

The metrics choice is task-specific. In Human-AI cooperative systems where the true data labels are known, it is possible to combine this information with the user decision to compute performance metrics (based on the confusion matrix), such as recall, FPR, precision, or false omission rate (FOR). These measures allow us to objectively quantify the impact of different components (*e.g.*, model score and/or different explanation types) in the human decision-making process. In practice, accuracy, recall, and FPR are better choices, because the denominator either depends on the sample size, or on the number of label positives and label negatives of the sample. Since these are constant over the course of the experiment and do not depend on the number of predicted positives and negatives (as it is the case for a metric such as precision and FOR), we can determine *a priori* the exact sample size for each metric.

In most systems, time is also a determining factor and should, therefore, be monitored during system modifications. In Human-AI decision making systems, explanations serve to help the human-in-the-loop to make a faster decision, by pointing them to what the model perceives to be the most important information for the decision. Consequently, this is an important aspect to measure when discerning the impact of explanations in decision-making processes.

Another relevant point, despite being more subjective, is the user's perception of the explanation quality, including its relevance and usefulness. For this reason, we propose a set of predefined five-point *Likert-type* scale questions, specified in Figure 2.

Finally, often times, decisions diverge from user to user. We expect the addition of more information (*e.g.*, model scores and/or explanations) to mitigate such differences. To accurately measure this effect, we use an agreement set where a subset of the data is shared between users with the intent of computing the metrics of agreement. We use *Fleiss' Kappa* [40] as the agreement metric because our experiments will incorporate multiple users. Additionally, we calculate the average agreement, which is the average pair-wise agreement between users.

3.2 Experimental Stages

While, in the first stage of the experiment, humans only have access to instance-specific information (feature data), in the second the human is provided with information of the model score, calibrated for simplification. Consequently, users may sometimes perceive it as a measure of *how confident the model is about predicting a given class*: scores closer to 1 or 0 express confidence, whereas scores around 0.5 convey more uncertainty.

The third stage of the experiment involves, in addition to the ML model score, the explanations. How and which information to show for which explainer should be defined in the experimental setup. There are many degrees of freedom when configuring an explainer: the explainer type (*e.g.*, self-explainable, *post-hoc*), the number of features to consider, how to represent the explanation (*e.g.*, feature contributions, heatmaps, scores, visualizations) as to minimize the cognitive load during the task execution. Another important aspect to pay attention to are the biases that may arise if explanation methods are distinguishable due to some factor (*e.g.*, their representation). Mitigating their representational differences is, therefore, a preventive step towards isolating the quality and relevance of the explanation methods from all the other possible visual factors.

3.3 Statistical Tests

The appropriate choice of a statistical test depends on two factors: (1) the metric distribution and (2) the end-goal of the test. Most statistical tests aim at identifying significant differences between measured averages of performance metrics in different scenarios (control *vs* treatment). In this case, we use of *Chi-squared* test [41] for multiple group comparison of instance-level binary metrics, such as accuracy, recall, and FPR. Conversely, for continuous performance metrics like decision time, we use a non-parametric test named *Kruskal-Wallis H* [42] to validate whether the samples belong to the same underlying process. This test is particularly suited for non-normal distribution of continuous variables.

We are interested in comparing pairs of groups and, specifically, in running comparisons between each variant and the control group. In these cases, we use *Chi-squared* test with the pairs to be tested in the performance values, and the Mann-Whitney U test [43] on continuous data. P-values must also be corrected for family-wise error rate with the *Holm-Bonferroni* method [44].

In order to quantify the perceived usefulness and relevance of the explanations measured through the questionnaire, we aim to identify distribution differences between different explainers for the proposed questions. We find the *Kruskal-Wallis H* to better suit

this goal when comparing multiple variants. To report the results of paired tests, we apply the *Kolmogorov-Smirnov* test [45] corrected with the *Holm-Bonferroni* method [46].

4 EXPERIMENTS

We employ our proposed application-grounded methodology, XAI Test, to evaluate and compare different explanation methods in a real-world decision-making task: fraud detection in payment transactions. We had access to a real fraud prevention system comprising a deployed ML model that predicts the risk of fraud for each payment transaction in a given online retailer. The fraud analyst is responsible for accepting or declining payment transactions for which the ML model is more uncertain about (the score is within a review band). This decision-making task is performed through a web interface in which the fraud analyst can inspect details of the payment transaction (*e.g.*, shipping address, billing email, time since last transaction) which represents the feature data (*i.e.*, *Data Only*) together with the risk score, given by the ML model, and an explanation.

While business requirements aim for more effective and efficient decisions, often, the model information is not sufficient to meet such criteria (*e.g.*, disagreement between fraud analysts and ML model or even mistrust in the model predictions). In an attempt to bridge this Human-AI gap, we conjecture that *explanations promote better human performance in such predictive fraud task*. Therefore, the prime goal of this experiment is to assess the real impact of showing explanations to real humans (the fraud analysts) interacting with a real ML model.

4.1 Experimental Hypotheses

As the first step of XAI Test, we formulated our hypotheses. Since we used a production system without permission to modify the ML model, we focus on the evaluation of *post-hoc* explanation[1] methods. With this in mind, we set out to answer the following hypotheses:

- *H1.* Showing fraud analysts the *ML Model Score* improves their performance [2] over *Data Only*;
- *H2.* Showing *post-hoc* explanations significantly improves human performance over *Data Only* and/or *Data + ML Model Score*;
- H3. Explanations from different *post-hoc* explainers impact humans differently; Assuming that humans trust the explanations, some explainers promote more effective and/or efficient decisions;
- *H4.* Each *post-hoc* explainer is perceived differently in terms of relevance, usefulness, and diversity;
- *H5.* Showing explanations increases fraud analysts agreement over the same set of transactions;
- *H6.* Showing model score information increases fraud analysts agreement over the same set of transactions.

[1]Explanations produced by post-hoc methods.
[2]Defined in Section 4.2.

4.2 Experimental Setup

We evaluate the above hypotheses using metrics indicative of the fraud analysts' performance in terms of both efficiency and efficacy at the decision-making task.

Metrics: We use the average decision time (of fraud analysts) as an efficiency measure and we use accuracy, FPR, and recall as measures of their effectiveness. Moreover, to address *H4*, we also measure their perceived relevance, usefulness, and diversity of the explanations through the questionnaire in Figure 2.

ML model: As an application-grounded evaluation of a real-world system, we used the fraud prevention system's ML model: a Random Forest's variant [47].

Explainers: Among the various XAI methods for tabular data, we opted for two of the most commonly used *post-hoc explainers*: LIME [9] and SHAP [3]. In particular, we leveraged the fact that the model is a decision tree ensemble to use the tree-based SHAP explainer - TreeSHAP [13]. We also included a third explainer specifically tailored for tree-based algorithms, known by ML practitioners as TreeInterpreter [48]. In terms of hyperparameters, we ran a few sensitivity tests to determine the most appropriate hyperparameters for the proposed task. From this analysis, we concluded that both SHAP and TreeInterpreter could be used with their out-of-the-box parametrization, whereas LIME had to be tweaked, specially, due to its stochastic nature[3]. Thus, besides the random seed, we also set the number of perturbed samples to 5k.

Explanation format: The explanations format for the three explainers consists of pairs of *feature-contribution*. We decided to only display the top 6 pairs based on contribution value. Unlike other tabular explanation formats, such as decision lists and decision sets [20]) the *feature-contribution* format benefits from its readability, simplicity and visualization flexibility.

Furthermore, to create a seamless experiment, we used this output's simplicity to homogenize the explanations representation across explainers. Given a set of feature-contribution pairs, we: (1) sort it in descending order by absolute contribution value[4], and (2) transform it into a human-readable format. This transformation comprises mapping the feature name to a natural language description plus parsing the feature value (*e.g.*, converting time from seconds to days).

We further added a color-based visual cue to reflect the changes in the associated suspicious risk (score): negative contributions represented with *green*, as they contribute for lower scores and consequently legitimate transactions, and, conversely, positive contributions represented with *red*. Figure 3, illustrates an explanation shown to a fraud analyst during the experiments.

Users: Three professional fraud analysts partook in the experiment. They were all experienced users of the fraud detection system used in the experiment.

Data: Two different samples were considered: (1) a training sample, derived from the same data set used to train the ML model, and (2) an experiment sample, from the production period of the ML model. We used the former as the background for LIME (to obtain information about features distributions). To create it, we

[3]LIME's internal local fidelity metric showed improvements exclusively upon variations on the number of perturbed samples.
[4]Higher contributions reflect more important features.

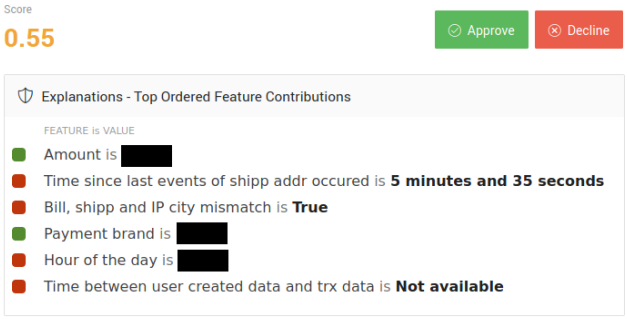


Figure 3: Visual representation of an explanation, as viewed by fraud analysts during the experiment (obfuscated to preserve privacy).

randomly sampled 100k transactions from the model's training set. Conversely, the sample for running the experiment itself, dubbed experiment sample, was extracted from the model's production period (November 2019), for which we had fraud labels. We extracted a stratified sample to attain 50% fraud prevalence.

To replicate a real scenario for the experiment the sample exclusively comprises transactions that lie in the review band, *i.e.*, transactions with higher model uncertainty. The final experiment sample size totals 1300 transactions. In the following section, we disclose how these transactions were distributed across the different experiment stages.

4.3 Experiment Outline

We conducted all three stages of the experiment, as XAI Test suggests (see Section 3.2). Given that each stage added levels of information, we decided to run them in a way that allows fraud analysts to incrementally stabilize their mental model of the task (as they adapt to new information within the system). This leads to the following experiment outline:

(1) *Data only*: information exclusively about the transaction (payment details and history) is available;
(2) *Data + ML Model Score*: both transaction data and the model score are available;
(3) *Data + ML Model Score + Explanations*: all of the above information is complemented with an explanation (from LIME, SHAP, or TreeInterpreter) of the model score.

As our baseline, we considered the stage where every information except the data was withheld from fraud analysts (*Data only* stage), as it allows us to isolate and quantify the real impacts of different information types in the human's performance in (and understanding of) the task. In the absence of prior knowledge about the metrics distribution for this particular task, we used a total of 400 transactions to conduct the two initial stages of the experiment (200 for each stage). Each of these samples were created without replacement from the experiment sample (stratified by fraud label to keep fraud prevalence at 50%). We found 200 transactions to be a good compromise between the pressing time and business constraints (*e.g.*, availability of the analysts) and the quality and rigor of the experiment.

On the other hand, we leveraged the results obtained in the initial experiments (*Data only* and *Data + ML Model Score*) to compute the sample size required to obtain significant results at the desired power, β, significance level, α, and effect size, δ. We set $\delta = 15$ because we found it to be a good compromise between sample size and the minimum difference detection. Moreover, we defined $\beta = 1 - \alpha = 0.9$, since we perceive both error types associated with statistical hypothesis testing (type I and type II) to be of equal importance during the experiment. In the end, and assuming the proxy estimates of the analysts' distribution were representative of their true performance, we concluded that a sample with 300 transactions would suffice for rigorously running the third stage of the experiment, the *Data + ML Model Score + Explanations* stage. Each sample was divided equally for each analyst. Each analyst reviewed the same number of transactions for every explainer in the experiment (100 transactions per explainer), which guaranteed the results were equally balanced and that the experiment results were not skewed towards a specific explainer or user.

To address hypothesis *H5*, we defined a subset of each sample to belong to an agreement set. In practice, this implies that all users reviewed the same exact transactions of the agreement set. This set accounted for about 12.5% of the transactions on every experiment stage.

5 RESULTS AND DISCUSSION

In this section, we evaluate how various levels of information affect the human's decision-making process in a fraud detection task. We first examine the impact of disclosing information about the ML model score when compared to withholding that information. We also analyse the impact of showing different *post-hoc* explanations on top of the information about the ML model score. We discuss the obtained results in terms of human effectiveness and efficiency at detecting fraudulent transactions.

Table 1 shows the experiment results for the conducted three-stage experiment (each stage reflects a group). Besides isolating the contributions of the different system components, this table also comprises the evaluation results of three popular *post-hoc* explanation methods, being one of the most comprehensive evaluation and comparison of XAI methods to date.

Our results show that data alone induces better decisions, while showing model scores or model scores with explanations significantly improves the decision time. Our results suggest that, in practical settings where decision speed is a main requirement, ML models explanations carry a significant speed up in human decision-making, as depicted in Figure 5. Additionally, data alone carries a better result in both accuracy and recall, registering even a significant difference in accuracy when compared to the group with model score, as depicted in Figure 4. Finally, we provide insights about the variability and agreement of the different *post-hoc* explainers based on the produced explanations for the experiment.

5.1 Data + ML Model Score

We first analyse the difference in between human decision-making with and without presence of the ML model score. We evaluate *H1*

Table 1: Performance, time and agreement metrics for all variants of the experiment. Statistical significance is tested between each explainer and each of the two groups that do not show explanations or only among explainers. ★ indicates significant difference with Data Only; no statistically significant difference was detected between each explainer and the Data + ML Model Score; ♦ indicates significant difference with all other explainers. The agreement metric is *Fleiss' Kappa*.

Group	Explainer	Sample Size	Metrics				Agreement
			accuracy (%)	recall (%)	FPR (%)	time (s)	
Data Only	-	200	**62.00**	**35.87**	15.74	59.50	0.41
Data + ML Model Score	-	200	51.02★	25.00	19.57	44.61★	-0.02
Data + ML Model Score + Explanations	LIME	300	58.59	27.03	**10.07**	50.29★	**0.53**
	TreeInterpreter	300	56.52	25.55	12.67	**43.03★♦**	0.30
	SHAP	300	59.73	31.08	12.00	45.72★	0.15

(see Section 4.1) in terms of the time taken to make decisions, accuracy, recall, and FPR, whereas *H6* is examined under the agreements measures mentioned in Section 3.1.

5.1.1 ***Showing end-users the ML Model Score improves average decision time over Data Only.*** Our results show that withholding the model score leads to significantly slower decisions. Using the *Mann-Whitney U* test, we detect a significant difference in times between *Data Only* and *Data + ML Model Score* ($p < 0.01$) (Table 1). A more thorough analysis of the performance metrics (see Figure 5) reveals an approximate decrease of 25% of the relative average time to decide, when presenting information about the model score. **When considering time as the performance metric, these results corroborate** *H1* (as defined in Section 4.1).

5.1.2 ***Showing end-users the ML Model Score deteriorates their accuracy over Data Only.*** Our results demonstrate that withholding information about the model score significantly improves the user's predictive accuracy. Table 1 shows that, after the application of the *Chi-squared* test, significant differences arise between *Data only* and *Data + ML Model Score* ($p = 0.08$). These results contradict *H1* (when using accuracy as the performance metric). This might derive from the fact that the instances being reviewed are in a score band near the decision threshold, and, therefore, have a higher associated uncertainty when being classified.

5.1.3 ***Showing end-users the ML Model Score does not significantly improve recall or FPR over Data Only.*** Our results do not exhibit statistically relevant improvements in terms of other users' performance metrics like recall or FPR. Considering these as the desired performance metric reveals to be inconclusive and, therefore, does not suffice to support nor reject *H1*.

In general, Figure 4 shows a degradation in all metrics derived from the confusion matrix, when comparing the *ML Model Score* group to *Data Only*, as both recall and accuracy registered a loss of 10% and FPR registered an increase of around 4% percentage points.

5.1.4 ***Showing the ML Model Score decreases agreement.*** The consensus among fraud analysts was shown to decrease as we incorporated more information. This is visible in Table 1, as the measurement of *Fleiss' Kappa* went from 0.41 in the *Data Only* variant to −0.02 in the *Data + ML Model Score* variant. The former reflects a setting where users, on average, agreed on the transaction label 76.67% of the times, whereas in the latter they only agreed on 63.33% of the times. This refutes the idea that showing more information would guide (or shape) users thinking process by giving hints about relevant aspects and, consequently, disproves hypothesis *H6*.

We hypothesize this large difference is due to (1) too small agreement set and (2) high proportion of transactions classified as legitimate (*i.e.*, 77%), leading to extra sensitivity to disagreements about fraudulent transactions.

5.2 Data + ML Model Score + Explanations

We further examine the performance differences between decision-making tasks involving *Data Only* and *Data + ML Model Score + Explanations*. In particular, we examine the impact of three distinct variants of the *Data + ML Model Score + Explanations* group: LIME, SHAP, and TreeInterpreter.

5.2.1 ***Showing post-hoc explanations significantly improves end-users average speed over Data Only.*** Figure 5 shows the confidence intervals of decision time for each group. By running a multiple group comparison using the *Kruskal-Wallis H* test, we observe statistically significant differences between explainer-based variants and the *Data Only* group ($p < 0.001$), which corroborates *H2* (when the performance metric is the reviewing time). We identify significant differences for every explainer, when they are compared pair-wise to the *Data Only* variant by using the *Holm-Bonferroni* corrected *Mann-Whitney U* tests we obtain p-values between 1×10^{-4} (for LIME) and 0.09 (for TreeInterpreter). When comparing against *Data + ML Model Score*, all explainers show increased decision time but this is not statistically significant.

5.2.2 ***Different post-hoc explainers impact the end-users decision speed differently.*** We also examine paired comparisons

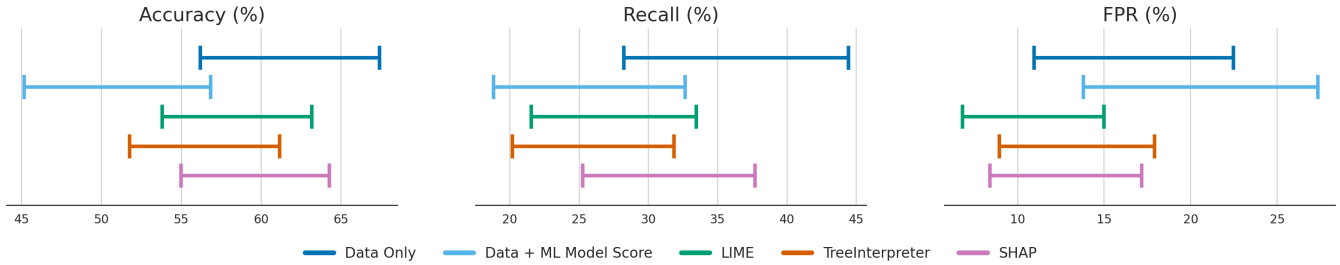


Figure 4: Confidence intervals (90%) for each performance metric of all variants of the experiment. The interval is calculated through the *beta distribution* for the estimated parameter p of each metric.

between the different explainers to address *H3* in terms of the decision efficiency. We detect significant differences when comparing LIME to TreeInterpreter ($p < 0.01$) and SHAP to TreeInterpreter ($p = 0.091$). In other words, results show that among the three evaluated *post-hoc* explainers, TreeInterpreter potentiates significantly faster decision-making processes. These results corroborate *H3* when considering the average time review as the fraud analysts' measure of performance.

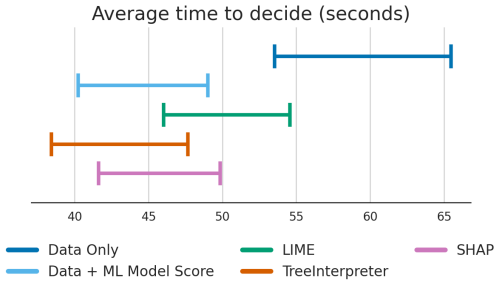


Figure 5: Confidence intervals for the average decision time of each variant. The interval represents the Standard Error of the sample multiplied by 1.64, representing a 90% Confidence Interval, centered around the mean group's mean.

5.2.3 ***Showing post-hoc explanations does not significantly improve end-users efficacy.*** In addition to efficiency, we examine the impacts of showing explanations to the human decision-making in terms of accuracy, FPR, and recall. As visible in Figure 4, all evaluated explainers are associated with deteriorated values for the predictive-accuracy metrics, except for the error-based metric, FPR. Effectively, although the values are not statistically significant, all explainers seem to lead to less false positives. Furthermore, as visible in Table 1 the multiple group comparison *Chi-squared* test provided no conclusive results and, consequently, no paired tests were conducted between the explainer variants. Notwithstanding the lower accuracy and recall values of each explainer when compared to the *Data Only* variant, explainers were still able to improve upon the results obtained for the *Data + ML Model Score* variant, although this improvement was also not statistically significant. The obtained results disprove *H2* and *H3*, when the considered performance metrics are either accuracy, FPR, or recall.

Notwithstanding these results, we emphasize that, performance-wise, the selected decision time metric is the most volatile metric and, therefore, the most susceptible to vary during the experiment due to some unaccounted external factors (such as connectivity issues or distractions).

5.2.4 ***Post-hoc explainers are perceived differently in terms of relevance, usefulness, and diversity by the end-users.*** We perform a multiple group comparison *Kruskal-Wallis H* test to compare the results obtained with the questionnaire in Figure 2. While no significant result is detected for the first question ($p = 0.238$), the test reveals significant changes relative to the second and third questions, that is, *"The explanation helped me review faster."* ($p < 0.001$) and *"The explanation was useful to help me make a decision."* ($p < 0.01$)). Figure 6 shows the distribution of the answers to the three questions posed during the last stage of the conducted experiment, discriminated by explainer. We observe that TreeInterpreter is the explainer with most positive answers (blue), especially in the third question. We also notice the high number of neutral answers, *neither*, and practically non-existing number of extreme answers, *i.e.*, *strongly agree* or *strongly disagree*. We can further observe, in statistical terms, that in the second question (middle), LIME registers a significant difference when compared to both SHAP ($p < 0.01$) and TreeInterpreter ($p < 0.01$). On the other hand, in the third question, no paired test registered a significant difference. In this question, TreeInterpreter is the explainer with results closer to significance. These results support *H4*, as each distinct explainers are indeed perceived differently by the users.

5.2.5 ***Showing explanations increases end-users agreement over the same set of transactions.*** We also examine the impacts in the agreement of the fraud analysts' decisions. Table 1 shows LIME to be the only explainer capable of improving the agreement beyond the *Data Only* group. However, when compared with the *Data + ML Model Score* variant, all explainer variants seem to evoke more consensus among the fraud analysts. Quantitatively speaking, LIME achieves by far the best agreement result with a *Fleiss' Kappa* of 0.53, and fraud analysts agree, on average, on 84.62% of the decisions. Also promising, but still inferior to the agreement achieved when all information is withheld from the user, is TreeInterpreter with a *Fleiss' Kappa* of 0.30, and with an average agreement of 69.23%. Lastly, SHAP exhibits the lowest value of agreement, with a *Fleiss' Kappa* of 0.15, an average agreement of 64.10%. These results

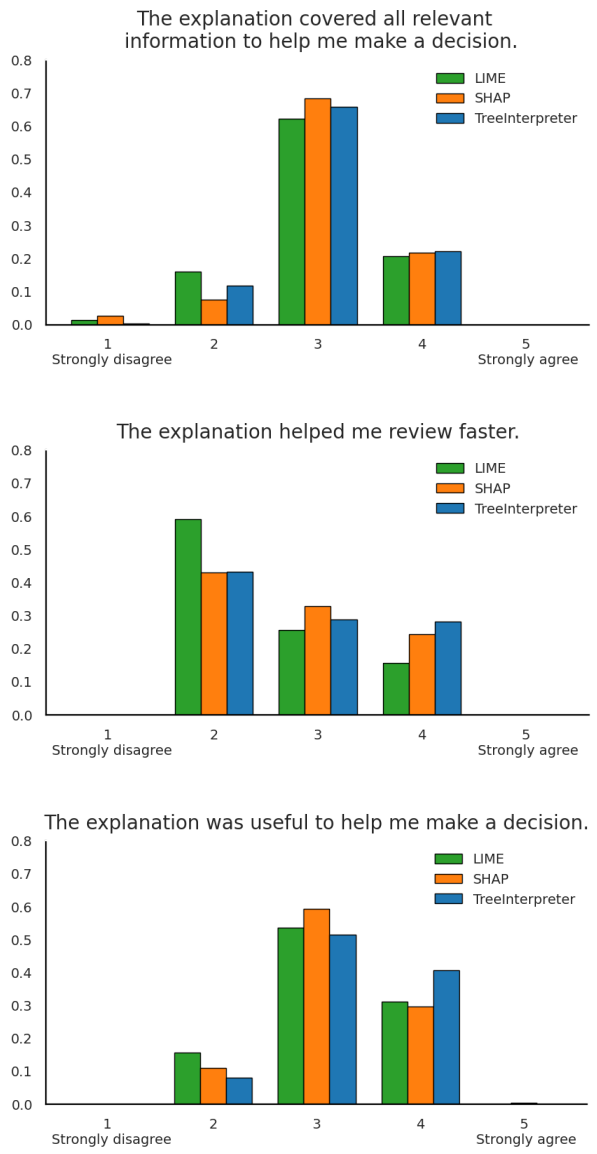


Figure 6: Distribution of answers to feedback questionnaire.

partially corroborate *H5*, as LIME actually seems to improve analysts agreement. However, the same does not verify for the other explanation methods.

5.3 Variability in Explanations

We analyse the variety and agreement of the explanations used during the third stage of the experiment. To this end, we collect the explanations of the different evaluated explainers (LIME, SHAP, and TreeInterpreter) for every transaction of the experiment. Each explanation comprises six feature-contribution pairs which are the basis of the explanations.

To better comprehend the explainers' behavior, we measured the diversity of their explanations. This implies comparing how many of the 111 available features are actually being used to create the explanations: LIME showed the least diversity, using 34 features (30.6% of the total set of features), followed by SHAP, which used a total of 89 features (80.2% of the total set of features), and TreeInterpreter, which used a total of 107 features (96.4% of the total set of features). Lower values in the number of used features translates into less variability in the explanations. This also ends up reflecting on the occurrence rate of the most popular feature (*i.e.*, the feature used the most times to explain an instance), which in LIME occurred in 89.7% of the transactions, as opposed to the most common feature in TreeInterpreter which only occurred 45.3% of the explanations.

The agreement between explainers is calculated by how many features two given explainers choose to integrate the explanation normalized by the length of the explanation. For example, if in an instance LIME and SHAP had chosen 2 features in common to explain the instance score, and the other 4 features were different for each explainer, the agreement in that instance would be 33.3% for that pair of explainers.

Comparing explanations between SHAP and TreeInterpreter produces an agreement of 53.0%, *i.e.*, 53.0% of the features used by SHAP for a given explanation were also used by TreeInterpreter. Likewise, the agreement for the other explainers' pairs produces an agreement of 41.0% (between LIME and SHAP) and 23.5% (between LIME and TreeInterpreter). These results show that the output explanation for a given instance depends on the *post-hoc* method chosen to explain it, *i.e.*, different explainers will choose different features to explain a given instance.

5.4 Study Limitations

In this section, we outline the main limitations of our empirical study. We have a constraint in the number of participants as well as their availability for the experiment, which in turn limits the sample size for the experiment. This has an impact on the effect size, or the rates of errors for the statistical tests. To perform tests with higher sensibility to smaller changes on the measured metrics, it is necessary to increase the sample size.

Another limitation of the study is that we cannot control all the possible external factors, such as difficulty of the instance, user attention to the tested information (data, model score, and explanations), connectivity speed, among other factors. However, the mitigation of the effects of such unaccountable factors is only possible when running large scale randomized controlled trials.

This study showed no significant differences in performance metrics derived from the confusion matrix between LIME, SHAP, and TreeInterpreter, using the same explanation format. A relevant study is to explore how different configurations and visualizations alter the observed results.

6 CONCLUSION

The recent developments of XAI methods has not been accompanied by a robust and practical assessment of their true impact on decision-making tasks. More often than not, the quality of these

methods is measured through proxy desiderata (*e.g.*, fidelity or robustness), hence, failing to convey the information of the actual impact on the end-users' performance (*e.g.*, accuracy or decision time). The lack of awareness towards the performance of the whole model + explanations + end-users may result in sub-optimal decision processes.

With this work, we hope to fill in this gap by proposing XAI Test, an application-grounded evaluation methodology suited for detaching the true impact of different information levels (*e.g.*, model score, explanations) in Human-AI collaborative systems. Following XAI Test, we conducted a user study to evaluate three well-known *post-hoc* explainability methods (i.e., LIME, SHAP, TreeInterpreter) on a real-world fraud detection task, encompassing 3 fraud analysts, an ML production model, and real-world data. Throughout the experiment, we progressively elevate the level of information presented to the analysts in three stages. We begin with information exclusively about the data (*Data only*) and subsequentially unveil information about the ML model score (*Data + ML Model Score*) and, in the last stage, about the explanations (*Data + ML Model Score + Explanations*). In the course of the experiment, we collect measures of the performance of the analysts in function of the revealed information. These include the duration, the accuracy, recall, and FPR of the decisions made, as well as the user's feedback on the perceived utility of the explanations.

To the best of our knowledge, this is the first study to perform a quantitative benchmark of the impact of different explanation methods on human decision-making performance on a real-world setting (real task, real data, real users). We complement this analysis with a strong battery of statistical tests to strengthen the validity of our conclusions. Obtained results reveal that, when provided with *Data only* information, fraud analysts decide significantly better but also more slowly when compared to variants that include information about the ML model. In this regard, our results show explanations (*Data + ML Model Score + Explanations*) to slightly improve the accuracy upon the *Data + ML Model Score* but to still fall short of the accuracy achieved in the *Data only* setup. Finally, amongst the three evaluated explainers, the analysts identify LIME as the least-favoured explanation method, potentially, due to its low explanations diversity.

In general, our results seem to suggest an existing trade-off between effectiveness and efficiency as the analysts are provided with added levels of information. This raises awareness towards blindly selecting popular *post-hoc* explanation methods in real-world decision-making settings.

7 ACKNOWLEDGEMENTS

The project CAMELOT (reference POCI-01-0247-FEDER-045915) leading to this work is co-financed by the ERDF - European Regional Development Fund through the Operational Program for Competitiveness and Internationalisation - COMPETE 2020, the North Portugal Regional Operational Program - NORTE 2020 and by the Portuguese Foundation for Science and Technology - FCT under the CMU Portugal international partnership.

REFERENCES

[1] General Data Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46. *Official Journal of the European Union (OJ)*, 59(1-88):294, 2016.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[5] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[6] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552*, 2018.

[7] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv*, pages arXiv–1811, 2018.

[8] Kasun Amarasinghe, Kit Rodolfa, Hemank Lamba, and Rayid Ghani. Explainable machine learning for public policy: Use cases, gaps, and research directions. *arXiv preprint arXiv:2010.14374*, 2020.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.

[10] Scott Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2, 01 2020.

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[12] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR.

[13] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[14] Gregory Plumb, Denali Molitor, and Ameet S. Talwalkar. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 2520–2529, 2018.

[15] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[16] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *CoRR*, abs/1802.00682, 2018.

[17] William W. Cohen and Yoram Singer. A simple, fast, and effective rule learner. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence and the Eleventh Innovative Applications of Artificial Intelligence Conference Innovative Applications of Artificial Intelligence*, AAAI '99/IAAI '99, pages 335–342, Menlo Park, CA, USA, 1999. American Association for Artificial Intelligence.

[18] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *Ann. Appl. Stat.*, 2(3):916–954, 09 2008.

[19] Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski. Maximum likelihood rule ensembles. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 224–231, 2008.

[20] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1675–1684. ACM, 2016.

[21] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N. Nair. Explainable neural networks based on additive index models. *CoRR*, abs/1806.01933, 2018.

[22] Xuezhou Zhang, Sarah Tan, Paul Koch, Yin Lou, Urszula Chajewska, and Rich Caruana. Axiomatic interpretability for multiclass additive models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 226–234, New York, NY, USA, 2019. ACM.

[23] Rich Caruana, Paul Koch, Yin Lou, Marc Sturm , Johannes Gehrke, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD'15, August 10-13, 2015, Sydney, NSW, Australia*. ACM, August 2015.

[24] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.
[25] Cynthia Rudin and Yaron Shaposhnik. Globally-consistent rule-based summary-explanations for machine learning models: Application to credit-risk evaluation. *SSRN Electronic Journal*, 01 2019.
[26] Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. Towards extracting faithful and descriptive representations of latent variable models. In *AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches*, 2015.
[27] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 303–310, 2018.
[28] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7775–7784. Curran Associates, Inc., 2018.
[29] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.
[30] Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, pages 3681–3688, 2019.
[31] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Jun Cai, James Wexler, Fernanda Viegas, and Rory Abbott Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). 2018.
[32] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
[33] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*, 2017.
[34] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, Oct 2019.
[35] Zachary C Lipton. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*, 2017.
[36] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.
[37] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. Case-based reasoning for assisting domain experts in processing fraud alerts of black-box machine learning models, 2019.
[38] Shi Feng and Jordan Boyd-Graber. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 229–239, New York, NY, USA, 2019. Association for Computing Machinery.
[39] Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 29–38, New York, NY, USA, 2019. Association for Computing Machinery.
[40] Joseph L. Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33(3):613–619, 1973.
[41] Karl Pearson F.R.S. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.
[42] William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, 1952.
[43] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60, 1947.
[44] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
[45] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.
[46] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
[47] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.
[48] Ando Saabas. treeinterpreter, 2015.