A Human-Grounded Evaluation of SHAP for Alert Processing

Hilde J.P. Weerts
h.j.p.weerts@student.tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

Werner van Ipenburg
werner.van.ipenburg@rabobank.nl
Rabobank
Zeist, The Netherlands

Mykola Pechenizkiy
m.pechnizkiy@tue.nl
Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

In the past years, many new explanation methods have been proposed to achieve interpretability of machine learning predictions. However, the utility of these methods in practical applications has not been researched extensively. In this paper we present the results of a human-grounded evaluation of SHAP, an explanation method that has been well-received in the XAI and related communities. In particular, we study whether this local model-agnostic explanation method can be useful for real human domain experts to assess the correctness of positive predictions, i.e. *alerts* generated by a classifier. We performed experimentation with three different groups of participants (159 in total), who had basic knowledge of explainable machine learning. We performed a qualitative analysis of recorded reflections of experiment participants performing alert processing with and without SHAP information. The results suggest that the SHAP explanations do impact the decision-making process, although the model's confidence score remains to be a leading source of evidence. We statistically test whether there is a significant difference in task utility metrics between tasks for which an explanation was available and tasks in which it was not provided. As opposed to common intuitions, we did not find a significant difference in alert processing performance when a SHAP explanation is available compared to when it is not.

KEYWORDS

explainable predictive analytics, human-grounded evaluation, human-computer interaction

Reference Format:
Hilde J.P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. In *Proceedings of KDD Workshop on Explainable AI (KDD-XAI '19).*

1 INTRODUCTION

Complex, black-box machine learning algorithms are increasingly applied in fields ranging from medicine to finance. In many contexts, the predictions are input for human decision-makers. Consequently, it is thought to be useful, if not crucial, to understand why a machine learning model made a prediction. However, as the complexity of the models increase, it becomes more difficult for humans to understand their behavior. To tackle this issue, recent efforts in explainable artificial intelligence (XAI) have resulted in many new explanation methods. However, the utility of these approaches in practical scenarios has not been researched extensively. Often, claims about interpretability and utility are based on proxies that have not been evaluated with real humans [1, 8]. As explanations need to be interpreted by real humans, a strong but solely theoretical foundation is no guarantee for utility.

In this work, we present a human-grounded evaluation to determine the utility of Shapley Additive Explanations (SHAP) for domain experts who assess the correctness of predictions, such as in medical diagnosis and fraud detection. In particular, we consider the utility for assessment of positive predictions, which we refer to as *alert processing*. SHAP is a state-of-the-art feature contribution method for explaining individual predictions [13, 18]. To determine the utility of SHAP, we perform two experiments in which real humans perform simplified alert processing tasks while alternately being provided with SHAP explanations.

Methods. Real humans performed simplified alert processing tasks, with and without an explanation of the model's prediction. Our approach is two-fold: (1) we statistically test whether there is a significant difference in task utility metrics between tasks for which an explanation was available and tasks in which it was not provided, and (2) we analyze the participants' written reasoning to determine the impact of different sources of evidence on the decision-making process, including the explanation.

Main findings. In contrast to common assumptions, we did not find a significant difference in alert processing performance between tasks for which a SHAP explanation was shown and tasks for which it was not shown. Our results suggest that possibly SHAP explanations alone are not that useful for alert processing. On the other hand, our qualitative analysis of the participants' reasoning during alert processing suggests that SHAP does affect the decision-making process. We speculate that possibly combining SHAP-based explanations with other techniques may provide higher utility for such tasks.

Outline. The present paper is structured as follows. In Section 2, we discuss related work on evaluating explanations of predictive modeling. In Section 3, we introduce several ways in which SHAP values may improve task utility as well as the corresponding hypotheses we formulated for the user study. Section 4 and 5 cover the experiment setup and results of the first and second experiment respectively. In Section 6 we present our concluding remarks.

2 RELATED WORK

Calling for more rigorous evaluations of XAI, Doshi-Velez and Kim [1] introduce a three-level taxonomy for evaluating explanations: *application-grounded*, *human-grounded*, and *functionally grounded*. *Application-grounded* evaluations consider the evaluation of real applications with expert users. *Human-grounded* evaluations consider real humans performing simplified tasks that either require or can benefit from interpretability. *functionally-grounded* evaluations use formal proxies of interpretability and do not require research with real human users.

Our study is an example of a human-grounded evaluation. Although only few works address this type of evaluations, we can identify several evaluation procedures that are commonly applied.

Output verification can be used to compare the interpretability of models [e.g. 6, 8, 9]. In this evaluation procedure, participants are asked to verify whether an output is consistent with the model. Another common approach for evaluating the interpretability of an explanation is *forward simulation* [e.g. 4, 6, 16, 18]. In forward simulations, it is determined how well humans can predict behavior of the model, after being exposed to an explanation [10]. Additionally, Doshi-Velez and Kim [1] suggest that different explanations can be evaluated by means of *binary forced choice*, in which humans review two alternative explanations of the same model and choose the best alternative.

An evaluation approach closely related to the present paper is *identification of incorrect behavior*. For example, Ribeiro et al. [17] study whether different explanations methods allow users to identify which classifier is likely to generalize to real world context. The capability of users to identify particular cases in which the model makes a wrong prediction under conditions with varying global interpretability was evaluated in [16]. In contrast to common assumptions, the authors found that exposing a model's global internals decreased people's ability to detect mistakes for unusual instances. This result shows the importance of user testing for validating intuitions on the utility of XAI.

SHAP explanations have been previously evaluated with human users in two ways. A forward simulation experiment, in which SHAP explanations significantly increased predictive performance, was reported in [18]. Other recent studies [e.g. 12, 13] show that among several explanation techniques, SHAP corresponds best with human intuitions of a simple decision tree model. Although these results provide evidence on the interpretability of SHAP, it does not directly follow that SHAP is useful for alert processing.

3 HYPOTHESES

In this section, we formulate research hypotheses on the utility of SHAP. To this end, we identify cases in which SHAP values might affect task performance for alert processing. We measure task performance in terms of *task effectiveness*, *task efficiency*, and *mental efficiency*, each of which could be impacted by providing a user with an explanation.

Task effectiveness refers to the extent to which the system helps the user to perform the task more effectively. For alert processing tasks, task effectiveness can be expressed as accuracy: the proportion of tasks in which the participant correctly distinguished between true positives and false positives.

Compared to just providing a model's confidence score, SHAP explanations could increase task effectiveness by increasing the ability of the user to assess model's credibility. For example, if a certain feature contributes substantially to the model's belief that an instance is positive, but the user assesses this reasoning as counter-intuitive, the user will be more likely to question the model's prediction. Contrarily but equally useful, the SHAP explanation may point the domain expert towards important feature values they would not have considered without being exposed to the explanation.

Hypothesis 1. *SHAP explanations increase task effectiveness of alert processing compared to the model's confidence scores alone, depending on the reasonableness of the explanation.*

Task efficiency refers to the extent to which the system helps the user to perform a task more efficiently, which we express as time spent on the task. Because SHAP explanations reveal which feature values are relevant for the model's decision, they can be used to determine whether the model's explanation is reasonable given domain knowledge. If it is, the domain expert might be able to process the instance more quickly.

Hypothesis 2. *If the number of features of an instance is sufficiently large, SHAP explanations increase task efficiency invested in alert processing compared to the model's confidence scores alone depending on the reasonableness of the explanation.*

Lastly, *mental efficiency* refers to the required mental resources to perform a task. Mental efficiency can be measured as self-reported mental effort [14]. According to cognitive load theory, the working memory has a limited capacity. For low-dimensional instances, SHAP explanations are unlikely to improve mental efficiency. Given the increase in information, they may even increase mental effort. On the other hand, a complex instance with many (tabular) features is unlikely to fit into working memory. In such cases, SHAP explanations may help the domain expert to focus on a subset of features that do fit into working memory, unless the explanation is unreasonable.

Hypothesis 3. *If the number of features of an instance is sufficiently large, SHAP explanations increase mental efficiency in alert processing compared to the model's confidence scores alone, depending on the reasonableness of the explanation.*

We test our hypotheses by means of two user experiments in which participants perform alert processing tasks while alternately being provided with SHAP explanations. In the first experiment, we measure the added value of SHAP explanations for each participant compared to the prediction probability alone. In the second experiment, we first explicitly quantify to what extent SHAP explanations agree with human intuitions. Subsequently, we measure the difference in task performance between participants who are provided with an explanation and those who are not. Moreover, we measure whether the difference depends on the extent to which the explanation aligns with human intuition.

4 EXPERIMENT 1: WITHIN-SUBJECT DESIGN

The first experiment is designed to measure the added value of SHAP explanations, compared to the model's confidence score alone. In order to mitigate user-specific effects, we adhere a within-subject experiment design.

4.1 Experiment Procedure

The first experiment consists of alert processing tasks followed by a written reflection. The alert processing tasks are performed in two rounds. The first round is designed for measuring mental efficiency, the second round for measuring task effectiveness.

Alert Processing Tasks. In each alert processing task, the participant is provided with an instance the classifier classified as positive. The participant is asked to predict the true class label and how much mental effort they invested to get to their answer. Each task belongs either to the *SHAP* or *NoSHAP* condition. In each task, participants are provided with the model's average confidence score (i.e. base value), the confidence score for the current instance, and the feature values of the instance. In the *SHAP* condition, the participants are also provided with a SHAP explanation (see Figure 1).

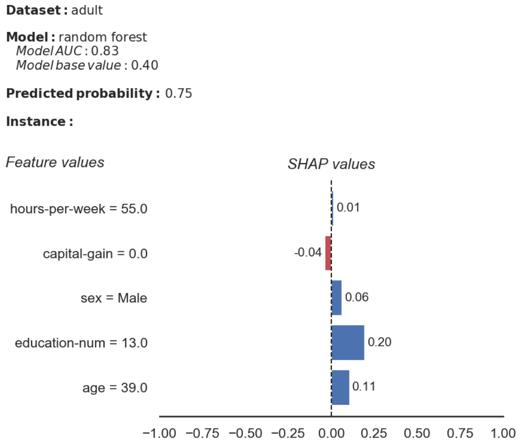

Figure 1: Example of an alert processing task in *SHAP* condition. In the *NoSHAP* condition, only the left part of the figure is shown.

Round 1: Mental Efficiency. Participants are provided with two sets of five instances, *A* and *B*. Each instance in set *A* is in the *NoSHAP* condition whereas each instance in set *B* is in the *SHAP* condition. The two sets are shown in order (see Figure 2).

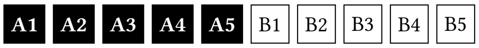

Figure 2: Setup of *mental efficiency* in Experiment 1. The letter (*A* or *B*) indicates the instance set, the number (1,2,3,4,5) the instance in the set. The color of the box indicates whether SHAP values are provided (white) or not (black).

Round 2: Task Effectiveness. Participants are provided with one set of ten instances the model predicted to be positives. Each instance is shown twice. The first time, an instance is shown in *NoSHAP* condition, the second time in the *SHAP* condition (see Figure 3). In this setup, any improvement or decrease in task performance will be due to additional information, which means that we can account for the difficulty of the instances. Note that this setup is not suitable for measuring the difference in mental effort, because the same instance is shown twice.

Participants' Written Reflections. After performing the alert processing tasks, participants discuss their results and experiences in small groups. In their written reflections, the groups discuss a.o. whether and how it would have been possible to distinguish between false positives and true positives for each of the ten instances of round 2, *task effectiveness.*

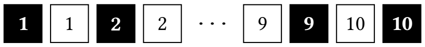

Figure 3: Setup of *task effectiveness* in Experiment 1. The number indicates the instance. The color of the box indicates whether SHAP values are provided (white) or not (black).

4.2 Experiment Details

Choosing an appropriate classification task for this user experiment is not trivial. On the one hand, the classification task should be non-trivial for humans. On the other hand, participants should have some domain knowledge about the data set to be able to argue about the reasonableness of an explanation. In the first experiment, we use the well-known *Adult* data set from the UCI repository [2] which we retrieved from OpenML [19]. The related classification task is to predict whether the income of a person exceeds $50,000 per year based on census data. We select five features and train a random forest classifier using the implementation of scikit-learn [15]. SHAP values were computed using the exact TreeSHAP algorithm proposed and implemented by Lundberg et al. [12].

A total of 102 students enrolled in an undergraduate introductory machine learning course participated in the first experiment.

4.3 Results of Experiment 1

To account for multiple testing and retain a family-wise error rate of $\alpha = 0.05$, we apply the Bonferroni correction. Accordingly, for each of the tests in the present work, we use a significance level of $\alpha = 0.005$.

Analysis of Task Effectiveness (Hypothesis 1).

Method. We measure task effectiveness through the proportion of correctly identified false positives and true positives. We test for a difference in the participants' accuracy before and after SHAP values are shown using McNemar's test. A post-hoc two one-sided equivalence test (TOST) procedure is used to assert whether the observed accuracy is equivalent [11]. We use an equivalence interval of $[-0.05, 0.05]$, i.e. the accuracy of participants within the two conditions is considered equivalent if the difference in accuracy is smaller than 0.05.

Results. We fail to reject the null hypothesis at a significance level of $\alpha = 0.005$ (χ^2(1, N=978) = 0.890, $p = 0.346$). Hence, it can be concluded that the difference in proportion of correct answers between *SHAP* (M=0.61, SD=0.49) and *NoSHAP* (M=0.59, SD=0.49) was not statistically significant. Moreover, the null hypothesis of the post-hoc equivalence test is rejected at $\alpha = 0.005$ ($z = -3.20, p_l =<$ .001, $z = 5.27, p_u =<$ .001). Hence, we can conclude that there did not exist a meaningful difference in accuracy between the two conditions.

Analysis of Mental Efficiency (Hypothesis 3).

Method. Mental efficiency is measured through self-reported mental effort, using the 9-point Likert-scale introduced by Paas [14]. As some instances may require much more mental effort than others, regardless of the *SHAP* condition, we first perform a one-way ANOVA to determine whether mental effort invested in particular tasks was significantly different from other tasks within either the *SHAP* or *NoSHAP* condition. These tasks are excluded from the remainder of the analysis. Subsequently, we test for difference in average mental effort spent in *SHAP* and *NoSHAP* by means of a two-sided paired t-test. A post-hoc TOST procedure using a one-sided paired t-test is used to assert whether the average mental effort in both conditions is equivalent, considering equivalence interval $[-0.5, 0.5]$.

Results. For samples related to the *NoSHAP* condition, none of the tasks had a significantly higher mental effort ($F(4, 404) = 1.89$, $p = 0.11$). For the *SHAP* condition, we do find a significant effect ($F(4, 404) = 19.02$, $p = 0.00$) and a multiple-comparison post-hoc analysis revealed that tasks 2 and 4 required significantly less mental effort than the other questions in the *SHAP* condition. Hence, data points related to these tasks are not considered in the paired t-test.

After asserting that the assumptions of normally distributed differences and the absence of outliers are met, we fail to reject the null hypothesis at $\alpha = 0.005$ ($t(100) = -0.66, p = 0.51$), which means that the difference in average mental effort in *SHAP* (M=4.81, SD=1.37, SE=0.14) and *NoSHAP* (M=4.74, SD=1.25, SE=-0.12) was not statistically significant. Moreover, the null hypothesis of the equivalence test is rejected ($t_l(100) = -4.37, p_l =< .001$, $t_u(100) = 5.68, p_u =< .001$). Hence, we can conclude that there does not exist a meaningful difference in invested mental effort.

Analysis of Recorded Participants' Reflections.

Method. A content analysis of the written reflections is performed by means of the grounded theory approach. Each of the reflection reports is coded with regard to the pieces of evidence that were used to make a decision about the true class of each of the instances.

Results. In total, 22 reports are analyzed. Ten types of evidence are identified, which can be further categorized into four main categories: SHAP values, feature values, the model's confidence score, and similar instances.

The primary source of evidence described in the written reflections is the instance itself, i.e. its feature values. After feature values, the model's confidence score is mentioned most often. SHAP explanations are used in three different ways. If an explanation is intuitive, participants see this as evidence of the correctness of the prediction. If an explanation is counter-intuitive, participants typically "adjust" the confidence score accordingly. For example, one of the groups argues that: "*the probability might be on the lower side but the SHAP values show that it is mainly brought down by having a positive capital gain which is quite counter-intuitive.*" In rare cases, participants change their initial beliefs based on the SHAP explanation. Finally, the true class of a similar instance is sometimes mentioned as evidence.

We would like to stress that the written reflections were performed in hindsight. Hence, the results do not necessarily reflect the participant's behavior *during* the alert processing tasks.

4.4 Conclusion from Experiment 1

There was no significant nor a practically relevant difference (larger than 0.05) in task accuracy before SHAP values were shown and after they were shown. Additionally, there was no significant nor meaningful difference (larger than 0.5) in average self-reported mental effort between instances for which SHAP values were provided and instances where SHAP values were not provided.

From the written reflections, we can conclude that apart from the instance's feature values, the leading source of evidence of the true class of an instance was the model's confidence score. This can be alarming, because raw confidence scores are often poorly calibrated; i.e., the predicted probabilities often doe not correspond to the true frequencies [7]. Consequently, confidence scores may be misleading for domain experts.

5 EXPERIMENT 2: CROSSOVER DESIGN

In the second experiment, we measure the difference in task utility metrics when SHAP values are shown compared to when they are not shown. Recall that Hypotheses 1, 2, and 3 include a notion of *reasonableness* of the SHAP explanation. In order to quantify the extent to which a SHAP explanation aligns with human intuition, the second experiment is preceded by a pretest experiment in which we measure human-assigned feature value contributions. The experiment setup is adapted from a within-subject to a crossover design.

5.1 Experiment Procedure

The experiment setup of the second experiment includes both a pretest experiment and a main experiment.

Pretest Experiment. In order to quantify to what extent SHAP explanations align with human intuitions, we ask participants to assign contributions to feature values of in total 20 instances. For each instance, participants are asked to explicitly indicate to what extent they believe a particular feature value would make it more unlikely, more likely, or would have no impact on the probability of belonging to the positive class. The participants are randomly assigned to two groups, group 1 and group 2. 10 of the instances are evaluated by group 1, the other 10 by group 2. After the data collection, the human-assigned contributions are compared to the corresponding SHAP explanations.

Main Experiment. In the main experiment, we adhere a cross-over design (see Figure 4). Each participant is randomly assigned to either group 1 or group 2. Two sets of instances are considered in the alert processing tasks, set *A* and set *B*. Group 1 will view instance set *A* in *SHAP* condition and set *B* in *NoSHAP* condition. Conversely, Group 2 will see set *A* in *NoSHAP* condition and set *B* in *SHAP* condition. In addition to the questions asked in the previous experiment, the participants are asked to provide their reasoning directly after each task; i.e. why they believe a particular instance is a false positive or true positive.

The crossover design has several advantages over the within-subject design of the previous experiment. First of all, in the previous experiment setup, participants had the option to change their answer after being exposed to a SHAP explanation. In the current setup, all information is always shown at once, which better resembles alert processing in a decision support scenario. Second, the new setup allows us to measure all hypotheses in the same experiment.

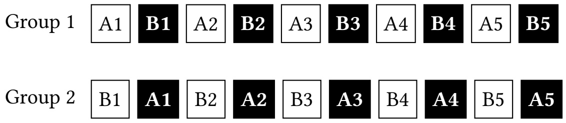

Figure 4: Setup of Experiment 2. The letter (A or B) indicates the instance set, the number (1,2,3,4,5) the instance in the set. The color of the box indicates whether SHAP values are provided (white) or not (black).

5.2 Experiment Details

In the second experiment, the UCI *Students Academic Performance* data set is used [5]. This data set contains student performance for mathematics in secondary education of two Portuguese schools. The associated classification task is to predict student's grades in mathematics based on a number of features including e.g. *age* and *number of previous failures*. We convert the regression task to a classification task, based on the minimum grade required to pass a course and retain the 13 most predictive features. Compared to the first experiment, the number of features is increased from five to thirteen. Similar to the previous experiment, we have split the data set into a training and test set and trained a random forest classifier.

A total of 20 undergraduate and graduate computer science or industrial engineering students participated in the pretest experiment. The main experiment was executed twice. In total, 57 people participated in the main experiment, consisting mainly of graduate and PhD students majoring in computer science or data science and having basic knowledge of XAI and SHAP.

5.3 Results of Experiment 2

Analysis of Agreement Between SHAP and Human Intuition (Pretest Experiment).

Method. For each instance, we quantify the agreement between human-assigned contributions and SHAP values as the average correlation. As SHAP explanations typically contain only a few large values and many smaller ones, it is desirable that higher weight is given to the top and bottom ranks. Hence we use a weighted signed rank correlation.

Results. The average SHAP agreement differs across instances but is typically not much lower than 0. For most instances, the model made a correct prediction and the corresponding SHAP explanations agree strongly with human intuitions ($\bar{R}_i > 0.20$).

Analysis of Task Effectiveness (Hypothesis 1).

Method. Given the crossover design, we require a more sophisticated statistical test than in the previous experiment. For Hypothesis 1, we use a generalized linear mixed model (GLMM) with a logit link function. We include a fixed effect for *SHAP* condition and agreement with human intuition. Following our hypothesis, we add an interaction effect between condition and agreement, and random effects for the participant and the alert processing task.

Results. Neither the coefficient corresponding to SHAP (M=0.12, 95% CI = [-0.71, 0.96], p = 0.76), agreement with human intuition (M=-7.17, 95% CI=$[-13.47, -0.87]$, $p = 0.026$), nor their interaction effect (M=0.70, 95% CI=$[-1.98, 3.38]$, $p = 0.61$) were statistically different from zero.

Analysis of Task Efficiency (Hypothesis 2).

Method. A linear mixed effects model is used to test for differences in speed. Again, we include the effect of *SHAP* condition, agreement with human intuition, and the interaction effect between *SHAP* and *agreement*.

Results. Even after data transformations, normality of the residuals could not be assumed. Hence, no conclusions can be made with regard to task efficiency.

Analysis of Mental Efficiency (Hypothesis 3).

Method. Mental efficiency is measured as self-reported mental effort. A linear mixed effects model is used, including the same effects as for the previous task performance metrics.

Results. The assumption of normally distributed residuals is reasonable. The SHAP main effect did not significantly differ from zero at $\alpha = 0.005$ (M=0.27, 95% CI = [-0.11, 0.65], p = 0.167), neither did the coefficient of rank agreement (M=-0.35, 95% CI = [-1.85, 1.15]) nor the interaction effect between SHAP and agreement (M=-0.94, 95% CI=[-1.16, 0.272], p=0.129).

Analysis of Recorded Participants' Reasoning. Recall that after each alert processing task, the participants are asked to articulate why they believed a certain instance was a false positive or a true positive.

Method. After several pre-processing steps including stop word removal and lemmatization, the replies are converted to vector-representation that indicates the presence of each term in each of the replies. For each combination of an instance and condition, the proportion of replies that contains a particular term is computed. Subsequently, for each of the instances, the proportions in the *SHAP* and *NoSHAP* conditions are compared. If the absolute percentage point difference between the two proportions is larger than 0.2, the corresponding replies are further inspected manually.

Results. In total, 20 terms were further inspected. It became clear that some of the differences were due to different wordings (e.g. *study time* versus *studytime*) and in some cases terms were mentioned because the participants did *not* agree with the SHAP explanation (e.g. "*I do not take into account that much the failures = 0*"). However, in five of the fourteen instances, the participants' reasoning did seem to be affected by the SHAP explanation (see Table 1).

In most of these cases, feature values of the instance are taken into account more heavily when presented with a relatively large SHAP value for that feature value (the largest absolute SHAP values in this data set typically ranged between 0.07 and 0.11). We have not identified any cases in which feature values that were taken

Table 1: Proportion of replies in which a feature was discussed in *NoSHAP* and *SHAP* condition.

Task ID	Feature	SHAP value	*NoSHAP*	*SHAP*
A2	*number of absences*	0.08	2/20	11/29
A4	*previous failure*	0.05	5/20	11/25
A7	*higher education*	-0.08	1/17	8/27
B3	*number of absences*	0.04	3/30	8/22
B6	*paid math classes*	-0.01	1/28	4/16

into account by people in the *NoSHAP* condition were not taken into account by participants in the *SHAP* condition.

5.4 Conclusion from Experiment 2

No significant differences in task effectiveness, task efficiency, and mental efficiency were measured when SHAP values were shown compared to when they were not available to the participants.

From the analysis of the textual replies, it can be concluded large SHAP values did affect the reasoning applied by our participants. These results suggest that large SHAP values can bring feature values of the instance to attention that are otherwise ignored.

6 CONCLUSIONS

XAI and related research communities have become productive in developing new interpretable machine learning methods. However, the evaluation of these methods often remains limited.

The results of the present paper suggest that it is important to perform evaluations with real users, rather than to rely on intuitions about utility. In neither of the two experiments it could be concluded whether SHAP explanations significantly impact task utility measured as task effectiveness and mental efficiency. Therefore, we cannot conclude that SHAP explanations are useful for human experts performing alert processing tasks. The post-hoc equivalence tests of our first experiment show that the failure to reject the null hypothesis was likely due to only a small difference in utility rather than a lack of data. This suggests that SHAP explanations alone are not that useful for alert processing.

Our analysis of the written reflections of participants of Experiment 1 has shown that, apart from the feature values themselves, the leading source of evidence was the model's confidence score. This is concerning, since confidence scores can be misleading.

Our textual analysis of the participants' reasoning in Experiment 2 has shown that large SHAP values can bring to attention feature values that are otherwise ignored. This shows that even though we could not identify a significant difference in task utility, the SHAP explanations did have an impact on the participants' decision-making process.

Future Work. We intend to pursue this direction further by performing a deeper analysis of the gathered data. It would be interesting to identify subgroups in the data for which the difference between *SHAP* and *NoSHAP* is exceptionally large. These subgroups could be described e.g. in terms of instance attributes and participant attributes. Such an approach could result in new hypotheses regarding the effect of local explanations on different aspects of alert processing performance. We intend to adopt an exceptional model mining approach introduced in [3] to automate this search.

Additionally, we would like to replicate the experiments with a classification task that contains a larger number of features and study how this affects task efficiency and mental efficiency.

REFERENCES

[1] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 http://arxiv.org/abs/1702.08608

[2] Dheeru Dua and Karra Taniskidou Efi. 2017. UCI Machine Learning Repository. Retrieved from http://archive.ics.uci.edu/ml.

[3] Wouter Duivesteijn, Tara Farzami, Thijs Putman, Evertjan Peer, Hilde J. P. Weerts, Jasper N. Adegeest, Gerson Foks, and Mykola Pechenizkiy. 2017. Have It Both Ways - From A/B Testing to A&B Testing with Exceptional Model Mining. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017*. 114–126.

[4] Allahyari Hiva and Lavesson Niklas. 2011. User-oriented Assessment of Classification Model Understandability. *Frontiers in Artificial Intelligence and Applications* 227 (2011), 11–19. https://doi.org/10.3233/978-1-60750-754-3-11

[5] Sadiq Hussain, Neama Abdulaziz Dahan, Fadl Mutaher Ba-Alwi, and Najoua Ribata. 2018. Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *J. Electrical Engineering and Computer Science* 9, 2 (Feb. 2018), 447. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

[6] Johan Huysmans, Karel Dejaeger, Christophe Mues, Jan Vanthienen, and Bart Baesens. 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems* 51, 1 (apr 2011), 141–154. https://doi.org/10.1016/J.DSS.2010.12.003

[7] Volodymyr Kuleshov and Percy S Liang. 2015. Calibrated Structured Prediction. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 3474–3482.

[8] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An Evaluation of the Human-Interpretability of Explanation. arXiv:1902.00006

[9] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1675–1684. https://doi.org/10.1145/2939672.2939874

[10] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. arXiv:arXiv:1606.03490

[11] Ying Lu and Judy A. Bean. 1995. On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Statistics in Medicine* 14, 16 (Aug. 1995), 1831–1839. https://doi.org/10.1002/sim.4780141611

[12] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. arXiv:1802.03888

[13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[14] Fred G. Paas. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology* 84, 4 (1992), 429–434.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[16] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. arXiv:1802.07810

[17] 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 SIG International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, New York, New York, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[18] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 3 (2014), 647–665.

[19] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. *SIGKDD Explorations* 15, 2 (2013), 49–60. https://doi.org/10.1145/2641190.2641198