Stop Ordering Machine Learning Algorithms by Their Explainability! An Empirical Investigation of the Tradeoff Between Performance and Explainability

Jonas Wanner, Lukas-Valentin Herm, Kai Heinrich, Christian Janiesch

Stop Ordering Machine Learning Algorithms by their Explainability! An Empirical Investigation of the Tradeoff between Performance and Explainability

Jonas Wanner[1] [0000-0002-0118-7757], Lukas-Valentin Herm[1] [0000-0002-0101-5429], Kai Heinrich[3] [0000-0002-4907-6802], and Christian Janiesch[1,2] [0000-0002-8050-123X]

[1] Julius-Maximilians-Universität Würzburg, Würzburg, Germany
{firstname.lastname}@uni-wuerzburg.de
[2] HAW Landshut, Landshut, Germany
[3] Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
kai.heinrich@ovgu.de

Abstract. *Numerous machine learning algorithms have been developed and applied in the field. Their application indicates that there seems to be a tradeoff between their model performance and explainability. That is, machine learning models with higher performance are often based on more complex algorithms and therefore lack interpretability or explainability and vice versa. The true extent of this tradeoff remains unclear while some theoretical assumptions exist. With our research, we aim to explore this gap empirically with a user study. Using four distinct datasets, we measured the tradeoff for five common machine learning algorithms. Our two-factor factorial design considers low-stake and high-stake applications as well as classification and regression problems. Our results differ from the widespread linear assumption and indicate that the tradeoff between model performance and model explainability is much less gradual when considering end user perception. Further, we found it to be situational. Hence, theory-based recommendations cannot be generalized across applications.*

Keywords: Machine Learning, Explainability, Performance, Tradeoff, User Study.

1 Introduction

Today, intelligent systems based on artificial intelligence (AI) technology primarily rely on machine learning (ML) algorithms [1]. Despite their prediction performance, there is a noticeable delay in the adoption of advanced ML algorithms based on deep learning or ensemble learning in practice [2]. That is, practitioners prefer simpler, shallow ML algorithms such as logistic regressions that exhibit a higher degree of explainability through their inherent interpretability [3].

In contrast, much of the current AI research focuses on the performance of ML models [4] and data competitions are dominated by deep learning algorithms such as artificial neural networks (ANN) that outperform shallow ML algorithms [e.g., 5]. However, the processing of these algorithms is practically untraceable due to its complex and

intransparent inner calculation logic. This renders it impossible for humans to interpret an ANN's decision-making process and prediction results, making it a black box.

This results in a tradeoff between performance and explainability which is not yet sufficiently understood. The uncertainty and lack of control due to a lack of explainability can fuel algorithm aversion of the end user. The aversion describes a phenomenon where users prefer humans over machines even when the performance of the machine is superior to the human [6]. In contrast, recent work by Logg, et al. [7] implies that for some situations when performance is communicated, humans may prefer machines resulting in algorithm appreciation. A better understanding of the tradeoff can help to reduce algorithm aversion and may even foster algorithm appreciation from an end user perspective.

While the performance of an algorithm can be estimated by common performance indicators such as precision, recall, or the F-score, it remains unclear, which ML algorithm's inherent interpretability is perceived as more explainable by end users. However, this is crucial as the perceived explainability of a prediction determines the effectiveness of an intelligent system. That is, if the human decision maker can interpret the behavior of an underlying ML model, he or she is more willing to act based on it [8] – especially in cases where the recommendation does not conform to his or her own expectations. As a consequence, intelligent systems without sufficient explainability may even be inefficacious as end users will disregard their advice.

In scholarly literature, several theoretical considerations on the tradeoff of performance and explainability exist [9-15], yet a scientific investigation or even an empirical proof is still missing. We formulate our research question accordingly:

"How do machine learning models compare empirically in the tradeoff between their performance and their explainability as perceived by end users?"

These insights have a high potential to better explain AI adoption of different ML algorithms contributing to a better understanding of AI decision-making and the future of work using hybrid intelligence. That is on the one hand, the results can help us to understand to what extent various ML algorithms differ in their perceived explainability from an end user perspective. This allows us to draw conclusions about their future improvement as well as about their suitability for a given situation in practice. On the other hand, the results can help us to understand how much performance end users are willing to forfeit in favor of explainability. Ultimately, Rudin [3]'s call to avoid explaining black-box models in favor of using inherently interpretable white-box models could be better approached if the tradeoff was sufficiently understood from a social-technical perspective.

In the following, Section 2 introduces fundamentals of ML and the state-of-the-art of existing ML tradeoff schemes concerning model performance and model explainability. In Section 3, we describe our methodology before we outline preparatory work comprising the datasets and algorithms. The section also comprises the technical realization of the algorithms, the measurement for comparison, and the survey design. In Section 4, we discuss the results of the empirical comparison. We close by summarizing our results and pointing out limitations of our study in Section 5.

2 Fundamentals and Related Work

2.1 Machine Learning Algorithms

ML focuses on algorithms that are able to improve their performance through experience. That is, ML algorithms are able to find non-linear relationships and patterns in datasets without being explicitly programmed to do so [16]. The process of analytical modeling building to turn ML algorithms into concrete ML models for the use in intelligent systems is a four-step process comprising data input, feature extraction, model building, and model assessment [1].

Each ML algorithm has different strengths and weaknesses regarding their ability to process data. Many shallow ML algorithms require the feature selection of relevant attributes for model training. This task can be time-consuming if the dataset is high-dimensional, or the context is not well-known to the model engineer. Common shallow ML algorithms are linear regressions, decision trees, and support vector machines (SVM). ANNs with multiple hidden layers and advanced neurons for automatic representation learning provide a computation- and data-intensive alternative called deep learning [1]. These algorithms can master feature selection on increasingly complex data by themselves [17]. In consequence, their performance surpasses shallow ML models and even exhibits super-human performance in applications such as data-driven maintenance [e.g., 18]. On the downside, the resulting models have a nested, non-linear structure that is not interpretable for humans, and its results are difficult to reproduce.

In summary, while many shallow ML algorithms are considered interpretable and, thus, white boxes, deep learning algorithms tend to perform better but are considered to be intransparent and, thus, black boxes [19].

2.2 Interpretability and Explainability in Machine Learning

Explanations have the ability to fill the information gap between the intelligent system and its user similar to the situation in the principal-agent problem [2]. They are decisive for the efficacy of the system as the end user decides based on this information whether he or she integrates the recommendation into his or her own decision-making or not. The question of what constitutes explainability and how explanations should be presented to be of value to human users fuels an interdisciplinary research field in various disciplines, including philosophy, social science, psychology, computer science, and information systems.

From a technical point of view, explainability in intelligent systems is about two questions: the “how” question and the “why” question. The former is about global explainability, which provides answers to the ML algorithm’s internal processing [3, 9]. The latter is about local explainability, which answers the ex-post reasoning about a concrete recommendation by a ML model [9]. To form a common understanding for our research artifact, we define explainability as “the perceived quality of a given explanation by the user” [19].

In this context, as noted above many shallow ML models are considered to be white boxes that are interpretable per se [13]. In contrast, a black-box ML model is either far

too complicated for humans to understand or opaque for a reason and, therefore, equally hard to understand [3]. Consequently in this research, in line with Adadi and Berrada [19]'s argument we consider a model's explainability as its innate interpretability by end users not using any further augmentations.

2.3 Related Work on Machine Learning Tradeoffs

Considerations about the (hypothesized) tradeoff between model performance and model explainability have been the subject of discussion for some time. Originating from theoretical statistics, a distinction for different ML algorithms was first made between model interpretability and flexibility [15]. More recently, this changed towards a comparison between model accuracy and interpretability [e.g., 10, 13] or algorithmic accuracy and explainability [e.g., 9, 12]. However, all tradeoffs address the same compromise of an algorithm's performance versus the algorithm's degree of result traceability.

Overall, in the field many subjective classifications of this tradeoff exist [9-15]. These subjective classifications of the different authors show great similarities but also some dissimilarities. We summarize the related work and their classifications (left side) in Figure 1 illustrating a high conformity between all authors. The resulting Cartesian coordinate system (right side) shows five common ML algorithms ordered by their common performance (*y*-axis) and their assumed explainability (*x*-axis). Grey-box models (i.e., ex-post explainers) are only subject of few studies [e.g., 12, 14], hence we have not included them in our considerations.

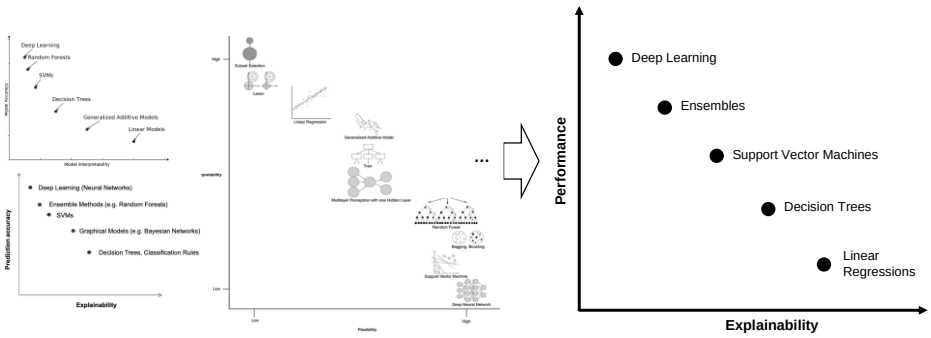

Fig. 1. A Synthesis of Common ML Algorithm Classification Schemes

While there is a general agreement on key ML algorithms, there are some differences on their placement and the granularity of representation. The general notion is that with a loss of performance, algorithms provide better explainability in a more or less linear fashion. That is, deep learning algorithms or ANNs are categorized as the most powerful with the least degree of model explainability, followed by ensemble algorithms, which consist of multiple ML models. Third in performance, SVMs serve as a large margin classifier based on data point vectors. Fourth, decision trees use sorted, aligned trees for the development of decision rules. Finally, linear regressions are considered of least performance, yet straightforward to interpret [20]. Some authors have chosen

to classify certain ML algorithms closer to each other to arguably represent better their assumed true position in the tradeoff [e.g., 9, 11, 21].

In essence, these theoretical classification schemes represent a hypothetical and data-centered view on the tradeoff of model accuracy vs. model interpretability. They have neither yet been validated for specific applications based on real data, nor with end users in a user-centered approach to unearth their true pertinency to represent said tradeoff of performance vs. explainability. Despite this obvious deficiency, they are commonly referenced as a motivation for user- or organization-centered XAI research or intelligent system deployment [e.g., 3, 21, 22].

Thus, in summary it remains unclear how the end users perceive explainability and how this is in line with these tradeoff considerations. In our work, we focus on the tradeoff between performance and an ML models inherent explainability to avoid biases introduced by model transfer techniques from the field of explainable AI (XAI), which aims at providing more transparent ML models that have both, high model performance and high explanatory power [11].

3 Methodology

Our research methodology uses four main steps: research question, data collection, data analysis, and result interpretation [23]. They are depicted in Figure 2.

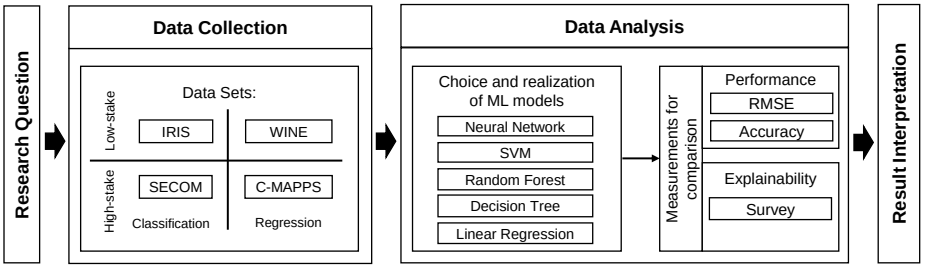

Fig. 2. Overall Methodology.

We started by formulating our RQ with the aim to shed light on the similarities and differences between different ML algorithms in terms of their tradeoff between performance and explainability. We verified the relevance of our RQ by a theoretical review of existing contributions and pointed out the research gap (cf. Section 2.3).

As we expect the tradeoff to be moderated by the underlying criticality of the task (low stake vs. high stake) and the type of the task (regression vs. classification), we employ a two-factor factorial design with four treatments using four different publicly available datasets. See Table 1 for an overview of the datasets.

To test the tradeoff empirically, we trained five ML models using common ML algorithms present in the aforementioned theoretical tradeoff schemes for the four treatment datasets using scikit-learn. We performed common data cleansing steps prior to model training. See Table 2 for an overview of the implementations.

Table 1. Overview of Datasets.

Dataset	Treatment	Description
IRIS [24]	Low-stake classification	IRIS is well known and has a low complexity. It contains 150 observations of 4 different features about the shape of iris flowers as well as their classification into one of 3 distinct species.
WINE [25]	Low-stake regression	The WINE quality dataset consists of 11 different features describing red Portuguese "vinho verde" wines. The dataset includes 1599 wine samples that are ranked in their quality from 0 to 12.
SECOM [26]	High-stake classification	SECOM includes use data from a semi-conductor manufacturing process. It contains data of over 590 sensors tracking 1567 observations of single production instances as well as the classification of semi-conductor production defects.
C-MAPSS [27]	High-stake regression	C-MAPSS provides turbofan engine degradation sensor data. It is based on a modular aero-propulsion system simulation about the remaining useful lifetime using different operational conditions. It contains simulation data from 93 turbines with 50 cycles per turbine and 25 sensors measurements per cycle.

To evaluate the performance of our models, we used two different measurements due to the type of problem (i.e., regression vs. classification). For the evaluation of the regression-based predictions, we applied the root mean square error (RMSE). For the evaluation of the classification-based predictions, we calculated the model's accuracy.

Table 2. Overview of ML Algorithm Implementations.

ML Algorithm	Implementation
Linear Regression	Due to data preprocessing, we skipped default normalization and used the default settings. For the non-centered datasets such as SECOM, we included the intercept of the model.
Decision Tree	We did not restrict the models by regulations such as the minimum sample split numbers of the estimators. The resulting trees have a depth of five or six, depending on the treatment.
SVM	For all datasets, we applied an SVM using a radial basis function as kernels.
Random Forest (ensemble)	We used the bagging algorithm random forest as proxy for ensembles. Random forests consist of 100 estimators each and their complexity was not restricted (see decision tree).
ANN	For C-MAPSS, we used an ANN with six alternating hidden layers consisting of LSTM and dropout layers. For the other datasets, we applied a multi-layer-perceptron with six hidden layers including dropout layers.

While a model's performance can be evaluated independently of the user, its explainability depends on the perceptions of its users [28]. Therefore, we evaluated the users' perceived explainability by conducting a survey to account for the subjective nature of the perception of the ML models. We used the platform prolific.co using a monetary incentive. We did not limit the participation by factors such as the experience with AI

or data science skills to receive broad feedback. For reasons of duration and repetitiveness, we designed two separate studies that were assigned at random: a classification study and a regression study, each containing a low-stake and high-stake case. The procedure within each variant was identical.

In the survey, we first collected demographics, prior experience with AI, as well as the participant's willingness to take risks. In the second part, we provided them with an introduction to the concepts of either regression- or classification-based ML, typical data processing steps, and general information about the visualization of ML predictions.

Second, we presented the use case for each treatment: The interviewees were asked to assume the role of an employee confronted with a decision situation. We provided a task definition and information about the process. Further, we explained that the task should now be performed by an intelligent system. For each case, we provided the criticality of wrong decisions.

Third, we evaluated their explainability based on the propositions by Hoffman, et al. [29]. To survey global explainability, we provided the participants with descriptions of the employed ML algorithms. To survey local explainability, we provided the participants with a graphical visualization of specific predictions. The participants did not receive any information about the performance of the ML to avoid biases. For each ML model, the participants had to rate their overall perceived explainability of the model on a five-point Likert-scale. The models were presented in random order to avoid sequence bias.

We received responses from 204 participants (112 classification, 92 regression). After processing multiple exclusion criteria (duration, lazy patterns, control questions), we could use 151 surveys (117 male, 34 female). Most participants (≈45%) were between 20 and 30 years old, followed by 31-40 (≈28%). ≈75% were from Europe, while ≈23% were from North America and only ≈2% from other regions. Half of the participants (≈52%) had no experience in AI, while ≈33% used AI for less than two years and only ≈15% had more than two years of experience with AI. ≈13% of the participants would describe their willingness to take risks as very low, while ≈46% would classify themselves as medium and ≈41% as high to very high.

4 Results

4.1 Result Comparison

Performance. In general, the performance results confirm the theoretical ordering in Figure 1 (*y*-axis). Nevertheless, the relative performance differs. Especially, the difference between random forest and SVM is smaller than assumed. In our case, this may be due to the datasets and the ensemble algorithm, but it reveals that the ordering of algorithms by their performance is hardly deterministic. Further, the performance difference between shallow ML algorithms and deep learning can be almost neglectable

in scenarios with low complexity such as IRIS. Still, linear regression constantly performed worst while ANN performed best in comparison to the other models. Table 3 illustrates the results of our performance evaluation.

Table 3. Performance Results of ML Models.

	Classification in Accuracy*		Regression in RMSE**	
Model	IRIS	SECOM	WINE	C-MAPSS
Linear Regression	81.59	68.70	1.05	59.39
Decision Tree	85.95	83.50	0.85	55.60
SVM	92.10	94.46	0.81	53.03
Random Forest	92.90	94.92	0.79	42.31
ANN	**94.21**	**95.20**	**0.77**	**38.56**

* higher = better, in %; ** lower = better, in total values

Explainability. We present the perceived level of explainability from the conducted survey for each algorithm in Table 4. We follow the recommendations of Boone and Boone [30] and applied a mean calculation for the Likert-scale data. The standard deviations appear normal with no discernible anomalies.

Table 4. Comparison of Mean Explainability and Standard Deviation.

	Mean Explainability*				SD Explainability**			
	Classification		Regression		Classification		Regression	
Model	IRIS	SECOM	WINE	C-MAPSS	IRIS	SECOM	WINE	C-MAPSS
Linear Regression	3.30	3.04	3.13	2.97	0.85	0.93	**0.86**	**0.85**
Decision Tree	**3.53**	3.34	**3.17**	**3.41**	**0.79**	0.83	0.88	0.90
SVM	3.29	3.12	2.88	3.03	0.96	0.89	0.90	0.85
Random Forest	3.38	**3.42**	3.32	3.32	0.91	**0.75**	0.87	0.90
ANN	3.07	3.25	2.92	2.95	1.02	1.01	1.00	0.98

* mean of five-point Likert scale; 1,00 = very low; 5,00 = very high; ** standard deviation of five-point Likert scale

Across all treatments, random forests and decision trees achieved the highest or second-highest ratings. Decision trees are considered highly interpretable by humans in terms of their global and local explainability, since it is possible to follow a path of variables from the root node to a leaf node containing the final decision [13]. This explainability by design makes the model itself (global) as well as every prediction (local) transparent. Random forests use multiple decision trees with a majority vote or averages on the predictions from the decision trees resulting in a single prediction. This could explain their comparably high scores. The perception of explainability varies across the remainder of models as discussed in the following.

4.2 Discussion

Low- and high-stake classification. For the low-stake classification treatment IRIS, the models' explainability were generally well-received and perceived as more similar. They reflect the theoretical ordering of explainability in Figure 1 (*x*-axis) quite well. IRIS represents a case of low algorithmic involvement with good accuracy values resulting in the low distances between the models. The case is straightforward with only few variables on flower properties such as sepal width. Hence, any participant should have been able to grasp the features relevant to fulfill this task in its entirety.

For the high-stake classification treatment SECOM, we found large performance differences as the case is more complex with more input variables, which is reflected by the poor performance of the shallow ML models such as linear regression. In addition, we found that the explainability of models, which can be visualized for simple cases in a straightforward way, loose their explanatory value for end users in this treatment.

We also found that the user's preference shifts from single decision trees to the majority vote of random forests. We assume that human biases may be at work more prominently in high-stake scenarios. This is also mirrored by the higher explainability scores of ANN for SECOM even though – objectively – the global and local explainability should be non-existent as ANN is a black-box model.

Low- and high-stake regression. The regression datasets also highlight the divergent perception regarding the different stakes. In the low-stake WINE treatment, the results mostly fit the theoretical assumption. In contrast, in the high-stake C-MAPSS treatment, the explainability score for ANN is higher than for SVM and linear regressions. Furthermore, linear regression received low scores for explainability in strong contrast to theory. A possible rationalization is the difficulty of the participants to grasp the nature of regression altogether since it is not as naturally understood as classification. This may highlight the importance of some data science skills at the human user's end in order for the explanations (also in the context of XAI) to have any meaningful impact on the (hybrid) decision-making.

In general, the random forest seems to master the tradeoff between performance and explainability particularly well in relative comparison to the other ML models. Except for decision trees, there is also a shift of the user's favor from shallow ML models to deep learning models when the stake rises.

Generalization of tradeoff. For the generalization of our findings and analysis of the tradeoff, we merged the data of the four treatments. In order to enable this merge, we normalized the data to the range of 0 to 1 to allow for relative comparison of the ML algorithms regarding the different use cases, tasks, and performance measurements. For the factor regression, we inverted the performance scale of RMSE since smaller values indicate better predictions, inversely to accuracy for classification. We transferred it into a Cartesian coordinate system similar to Figure 1. We used mean values to yield a position for each algorithm. Figure 3 shows the resulting averaged scheme calculated from the data in Tables 3 and 4.

The hypothetical simple linear relation between ML model performance and ML model explainability assumed theoretically by prior research does not hold across our

user-centered treatments. While we can confirm some tendencies mostly concerning ML model performance, reflected by accuracy and RSME, a few things are notably different from the theoretical proposition.

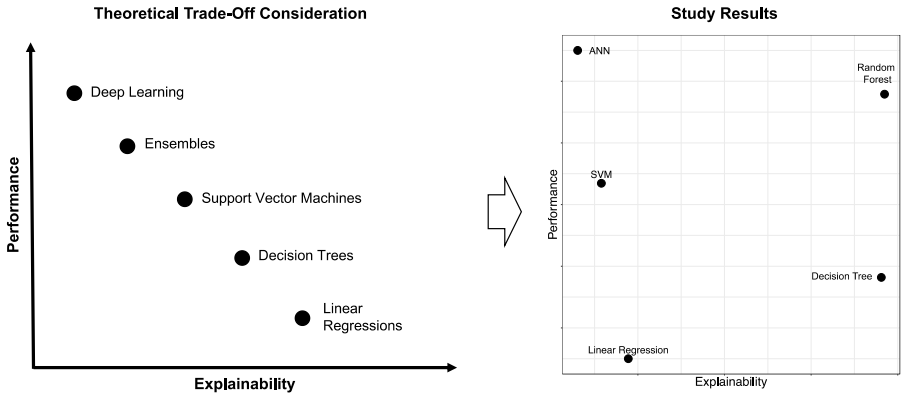

Fig. 3. Theoretical vs. Empirical Scheme for the Tradeoff of Performance vs. Perceived Explainability in Machine Learning

We find that the tree-based models decision trees and random forests are perceived to provide the best explainability of the five ML models by far from an end user's perspective. We assume that this is most likely due to their intuitive transparency with regard to global explainability [31], which may indicate that these two tree-based algorithms do not invoke the same degree of algorithm aversion associated with the remainder of ML algorithms. Contrary to our expectations, we could not substantiate that a single decision tree is perceived as more explainable than a random forest consisting of many unbalanced decision trees. We assume that this may be since we did not present all resulting trees of the random forest to the participants for review.

4.3 Implications

Our observations enable us to suggest theoretical and practical implication. They are important to consider when assessing how people respond to algorithmic advice as they hold implications for any decision maker or organization using intelligent systems.

Biases hinder objective measurement. It is possible that participants were biased in their judgement by the perceived capability or promise of an algorithm and therefore assumed a higher value [32]. That is, shallow ML algorithms such as SVM and linear regression offer a form of internal explainability. Hence, they were supposed to result in a better perceived explainability than black-box models with no internal explainability such as ANN. However, we found that there is hardly any difference in their perceived explainability by end users. This may be due to participants who were not able to understand the presentation of SVM and linear regression as they lacked prior knowledge [33], which may be a practical problem in real-life cases as well. In contrast,

simpler models seem to be especially good in explaining more straightforward scenarios. Consequently, due to high valuation in one category (performance), end users may attribute higher scores in another category (explainability). This is called halo effect.

Interpretability does not entail explainability. The discrepancy between theory and our empirical findings can be explained at least partly by the nature of our observations. While theoretical contributions look at the algorithmic and mathematical description of objects (data-centered perspective), we have employed a socio-technical and thus user-centered perspective. That is, in our study, we targeted the naturally biased perception of end users of an ML algorithm directly and found that the difference between performance and explainability is not linearly increasing. Rather, we found that linear regression's and SVM's (and ANN's) explanatory value is far from tree-based algorithms in most situations. While our results do not allow to uniformly rank and rate explainability for ML decisions (and were not expected to), they add to the growing evidence that there is more to model explainability than transparent mathematical parameters and good intentions. Moreover, ordering ML algorithms by their assumed data-centered interpretability is not helpful as it is constantly being misinterpreted and misused in socio-technical settings. In contrast, socio-technical aspects stand out as important for the efficacious use of ML models and explainability may be the key factor for their acceptance by end users [34, 35]. According to our research, in non-augmented form decision trees and random forests are currently the most suitable options to engage with end users.

5 Conclusion, Limitations, and Outlook

Albeit its fundamental importance for human decision-makers, empirical evidence regarding the tradeoff between ML model performance and explainability is scarce. The goal of our research was to conduct an empirical study to determine a more realistic depiction of this relationship and subsequently compare the placement of common ML models to the existing theoretical propositions.

We found that the explanatory value of decision trees and random forests constantly dominates other ML models. Comparing averages, we could not find noteworthy differences in the perceptions of explainability of SVM, linear regression, and even ANN. We did notice though that explainability was generally better received for more straightforward cases such as low-stake classifications.

In summary, we found existing theoretical propositions to be data-centered and misleading oversimplifications when compared to our user-centered observations. Our study shows that when explanations are put to use, socio-technical factors of user perception dominate well-intended analytical considerations concerning the goodness of visualizations by ML experts.

As with any empirical research, our study faces some limitations. First, our study was an online survey with benchmarking datasets. While we only allowed for participants with a certain background, participants may have been exposed to the scenarios and several of the ML algorithms for the first time. Hence, we measured an *initial* ex-

plainability. Second, there was no time restriction for viewing and assessing an explanation. We expect results to differ in a high-velocity treatment. Third and last, we compared inherently interpretable shallow ML algorithms and ANN without further augmentations. We assume that XAI augmentations will affect explainability positively. In contrast, other more diverse ensembles than random forests may perform worse.

Concluding, we identified socio-technical aspects as highly important for the perception of explainability and therefore further user studies with varying skill levels and cultural backgrounds are necessary to better understand the biases at work. Further, explainability does not entail understandability. If explainability only contributes to more trusted decision-making but not to a better understanding, research into XAI may be on the wrong track and ultimately only lulls users into a false sense of security by adding fancy yet inefficacious visualization.

References

1. Janiesch, C., Zschech, P., Heinrich, K.: Machine Learning and Deep Learning. Electronic Markets forthcoming, (2021)
2. Wanner, J., Heinrich, K., Janiesch, C., Zschech, P.: How Much AI Do You Require? Decision Factors for Adopting AI Technology. In: Proceedings of the 41st International Conference on Information Systems (ICIS), pp. 1-17. AIS, India (2020)
3. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206-215 (2019)
4. La Cava, W., Williams, H., Fu, W., Moore, J.H.: Evaluating recommender systems for AI-driven data science. arXiv:1905.09205 (2019)
5. Hyndman, R.J.: A brief history of forecasting competitions. International Journal of Forecasting 36, 7-14 (2020)
6. Burton, J.W., Stein, M.K., Jensen, T.B.: A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making 33, 220-239 (2019)
7. Logg, J.M., Minson, J.A., Moore, D.A.: Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes 151, 90-103 (2019)
8. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1135-1144. San Francisco, CA (2016)
9. Dam, H.K., Tran, T., Ghose, A.: Explainable software analytics. In: Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER), pp. 53-56. Gothenburg (2018)
10. Yang, Y.J., Bang, C.S.: Application of artificial intelligence in gastroenterology. World journal of gastroenterology 25, 1666–1683 (2019)
11. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web 2, (2017)
12. Angelov, P., Soares, E.: Towards Explainable Deep Neural Networks (xDNN). arXiv:1912.02523 (2019)
13. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82-115 (2020)

14. Nanayakkara, S., Fogarty, S., Tremeer, M., Ross, K., Richards, B., Bergmeir, C., Xu, S., Stub, D., Smith, K., Tacey, M.: Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study. PLoS medicine 15, e1002709 (2018)
15. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning. Springer, New York (2013)
16. Bishop, C.M.: Pattern recognition and machine learning. Springer, New York (2006)
17. Schmidhuber, J.: Deep learning in neural networks: An overview. Neural networks 61, 85-117 (2015)
18. Wang, J., Ma, Y., Zhang, L., Gao, R.X., Wu, D.: Deep learning for smart manufacturing: Methods and applications. Journal of Manufacturing Systems 48, 144-156 (2018)
19. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138-52160 (2018)
20. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press Cambridge (2016)
21. Guo, M., Zhang, Q., Liao, X., Chen, Y.: An interpretable machine learning framework for modelling human decision behavior. arXiv:1906.01233 (2019)
22. Asatiani, A., Malo, P., Nagbøl, P.R., Penttinen, E., Rinta-Kahila, T., Salovaara, A.: Sociotechnical Envelopment of Artificial Intelligence: An Approach to Organizational Deployment of Inscrutable Artificial Intelligence Systems. Journal of the Association for Information Systems 22, 325-352 (2021)
23. Müller, O., Junglas, I., Brocke, J.v., Debortoli, S.: Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems 25, 289-302 (2017)
24. Marshal, M.: Iris Data Set. (1988), https://archive.ics.uci.edu/ml/datasets/iris
25. Cortez, P.: Viticulture Commission of the Vinho Verde Region (CVRVV). (2009), archive.ics.uci.edu/ml/datasets/wine+quality
26. McCann Michael, J.A.: SECOM Data Set (2008), archive.ics.uci.edu/ml/datasets/secom
27. Saxena, A., Goebel, K.: Turbofan engine degradation simulation data set - NASA Ames Prognostics Data Repository. (2008), www.ti.arc.nasa.gov/tech/prognostic-data-repository/#turbofan
28. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 267, 1-38 (2019)
29. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608 1-50 (2018)
30. Boone, H.N., Boone, D.A.: Analyzing likert data. Journal of extension 50, 1-5 (2012)
31. Mohseni, S., Zarei, N., Ragan, E.D.: A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv arXiv: 1811.11839 (2018)
32. Hilton, D.: Mental models and causal explanation: Judgements of probable cause and explanatory relevance. Thinking & Reasoning 2, 273-308 (1996)
33. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K.: Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1-13. ACM, Glasgow (2019)
34. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-adapted interaction 18, 455-496 (2008)
35. Lee, M.K.: Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society 5, 205395171875668 (2018)