Technical Challenges and Opportunities in Explainable Artificial Intelligence: A Survey

Willem van der Spek
Vrije Universiteit Amsterdam, The Netherlands
Netherlands eScience Center, The Netherlands
w.vanderspek@esciencenter.nl

ABSTRACT

Recent advancement of artificial intelligence (AI) in the past few years have increased the complexity of models far beyond the bounds of human intelligibility. In order for AI to yield the best of expectations across its many application fields, the barrier of explainability has to be addressed. Paradigms underlying this problem constitute the eXplainable AI (XAI) field, which is deemed to play a key role in order to further adopt AI. With XAI being as relevant as ever, this work aims to provide an overview of the field through summarizing previous efforts, gathered through systematic literature search and analysis. This work has further contributed by formalising the main technical obstacles the field is facing and highlighting how other works have contributed to either expose or attempt to refute these challenges. Our prospects have led to a clear outline of the field, where we hope our work to serve as reference material to stimulate further research.

KEYWORDS

Systematic Literature Review, XAI, Explainable AI, Interpretable Machine Learning, IML, Machine Learning, ML

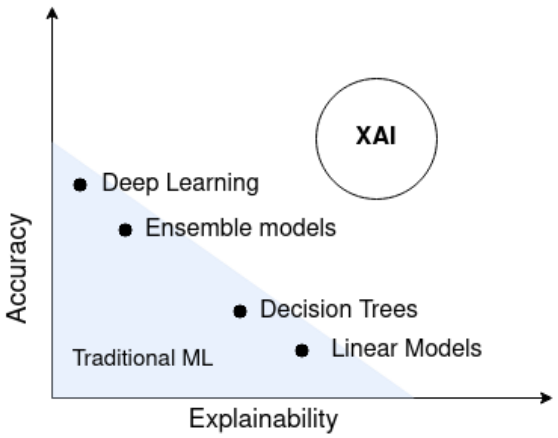


Figure 1: The accuracy versus explainability trade-off. Traditional Machine Learning models are confined within the blue area, whereas the goal and vision of XAI is create models that are both explainable and accurate.

1 INTRODUCTION

In the past decades, research trends in computer science have been increasingly moving towards covering the domain of Artificial Intelligence (AI) and more specifically Machine Learning (ML) and Deep Learning (DL). The predictive models that have been enabled through these fields have similarly vastly increased in complexity in order to maximise their predictive power [10].

Nevertheless, this focus on strongly prioritising the model's accuracy above all else has gained increased criticism as it generates *black-box models* that are inherently non-transparent in their decision-making preventing users from properly assessing, understanding and possibly correcting the models [1, 10, 30]. In fact, colossal traction is gathering on imposing a six-month moratorium on the development of any Large Language Model (LLM) more complex than *GPT-4* [26]. These concerns become especially stringent as AI is moving towards mission-critical domains, where AI-driven decisions can have a profound impact on human lives, such as medical imaging [20] and criminal justice [16].

As a result of the ever-growing concerns regarding black-box models, the field of *eXplainable Artificial Intelligence* (XAI) is dedicated to addressing the issue of non-transparency with the goal and vision of transparent, fair and accountable models. The field has been steadily increasing in size, which we see reflected in several scientific events. Some examples include the conference on Fairness Accountability and Transparency (FAccT) and the workshop on Human Interpretability in Machine Learning (WHI) hosted by the International Conference on Machine Learning (ICML) where XAI plays a key role. Even though the advances in XAI have been successful in making models more transparent and socially acceptable, the field has not yet reached full maturation, hence XAI is still facing some hurdles in its way towards adoption and gaining trust in a variety of communities [9].

Motivated by the visions of XAI and these obstacles, this work aims to capture field of XAI as a whole by reviewing some of its concepts and taxonomies. Furthermore, we formalise challenges in the field of XAI and their relevant contributions. We will also present our method to identify relevant research.

2 RELATED WORK

Several literature reviews on XAI, attempting to describe the field through its multiple related disciplines, have been published in he past few years. An instrumental work was done by Barredo Arrieta et al. [9] where the authors made several key contributions. Firstly, a novel definition of the *target audience* as a key element of explainability was proposed, which will be further discussed in section 5. In addition, an outline of challenges and future research directions were given. Contrary to our work, these challenges do not only focus on the models per se, but also the model development process and . Our work aims to focus on the contributions provided

by the predictive modeling community and identify some technical challenges.

the work of Vilone and Longo [40] aims to provide some of the recent advancements in the field and capture its main concepts by introducing a large set of taxonomies and classifications in order to effectively capture relevant research trends. Similarly, Das and Rad [12] provide a taxonomy of the field as well as an extensive set of *interpretable machine learning* (IML) methods. The authors also list some core challenges and identify new research opportunities to follow. Similarly to our work, the apparent lack of evaluation methods, sensitivity to adversarial attacks and limitations of explanation map visualisations were identified as main challenges for XAI. Contrary to these works, our work mainly focused on the challenges and has formalized these challenges by summarizing a main body of challenges in table 1 identified through relevant literature. We have proceeded to provide intuitions and insights regarding these challenges and share the recent advancements proposed to address them.

3 STUDY DESIGN

3.1 Research Goal

With this study we aim to create an overview of the field of the XAI and identify its challenges and opportunities for further research. Specifically, we aim to review the technical challenges in the field. With that goal in mind, we have defined a set of research questions to properly identify the scope of this project. We have gathered a body of relevant works that aim to answer these questions.

3.2 Research Questions

Following our earlier defined goal, we aim to concretely achieve it through answering the following Research Questions (RQs):

RQ1 What are the main challenges that XAI is facing currently?
RQ2 Are there studies suggesting that XAI methods have predictable behaviour?
RQ3 Is there a methodology for ensuring the reliability of XAI methods?

3.3 Initial Search

Organising the literature of explainable AI within a single review within reasonable scope is a non-trivial task. Both the concept of XAI and its applications are strongly multidisciplinary [9, 40]. As a result, we opted to exclude the following works:

1) Works solely focused outside of the field of predictive modeling. For instance, some of the works found focused on improving XAI through user-grounded evaluations, which were excluded.
2) Studies focusing on the application of XAI algorithms to specific problems, rather than expanding the field. For instance, papers focused on applying XAI in the medical imaging domain were discarded.
3) Works not available in English.

We proceeded to conduct search queries within the ACM digital library, IEEE Xplore, Scopus and Google Scholar. The following terms were used to find papers: *'explainable artificial intelligence'*, *'interpretable machine learning'*, *'explainable machine learning'*, *'XAI'*, *'IML'*. It was noted that term Interpretable Machine Learning (IML) is similar to eXplainable AI (XAI). Although some works propose that there is a nuance between the two terms, our work has solely used the term XAI for the sake of consistency [1]. In order to gain access to a wide variety of conferences and journals, we opted to diversify our digital libraries and employed these search queries on *ACM digital library*, *IEEE Xplore*, *Scopus* and *Google Scholar*.

3.4 Further analysis

It was not feasible to properly analyse all of the results from the systematic search given that the sheer volume of papers would be too large and out-of-scope (numbering over a thousand works) for our review. Instead, we identified a smaller body of *main works* [9, 13, 22, 24, 31–33, 35, 39, 40] and employed a snowballing approach to further identify relevant works. We set an additional requirement for these works to be peer-reviewed. Two other works outside of scientific literature bearing significant relevance to this research were further added and archived [25, 26], this yielded a total of 43 works.

4 OVERVIEW OF THE ALGORITHMIC LANDSCAPE

In the field of XAI, a variety of taxonomies have been proposed to properly identify each algorithm in the current landscape [1, 9, 12, 24, 39, 40]. The XAI field has grown to such proportions that a single taxonomy might not be sufficient to properly conceptualize it. Nevertheless, recent reviews of these taxonomies have suggested a de facto standard which has been presented in Figure 2 [39]. The dimensions of the taxonomy could be summarized as follows:

1) **Scope**: Regarding their scope, explanations are usually divided between *local* and *global* scope. Local scope refers to a explaining a single instance in the dataset, for instance a prediction result from a single image. Global explanations, on the other hand aim to explain the model on the fully aggregated data, e.g. finding the importance of features on the model.
2) **Stage**: XAI Algorithms are divided by the stage in which explainability is seeded into the model. *Ante-hoc* and *post-hoc* mean that the explainability is introduced prior and after the training phase of the model respectively.
   (a) **Applicability** For post-hoc methods, the further distinction is made between *model-specific* and *model-agnostic* methods. The former comprises algorithms which inner workings depend on both the internals of the model and the architecture of the model. The latter has no such constraints and simply uses inference of the black-box model to generate explanations.

   Marked examples of post-hoc model-agnostic methods include *Local Interpretable Model Agnostic Explanation* (LIME) [35] and *SHapley Additive exPlanation* (SHAP) [31].

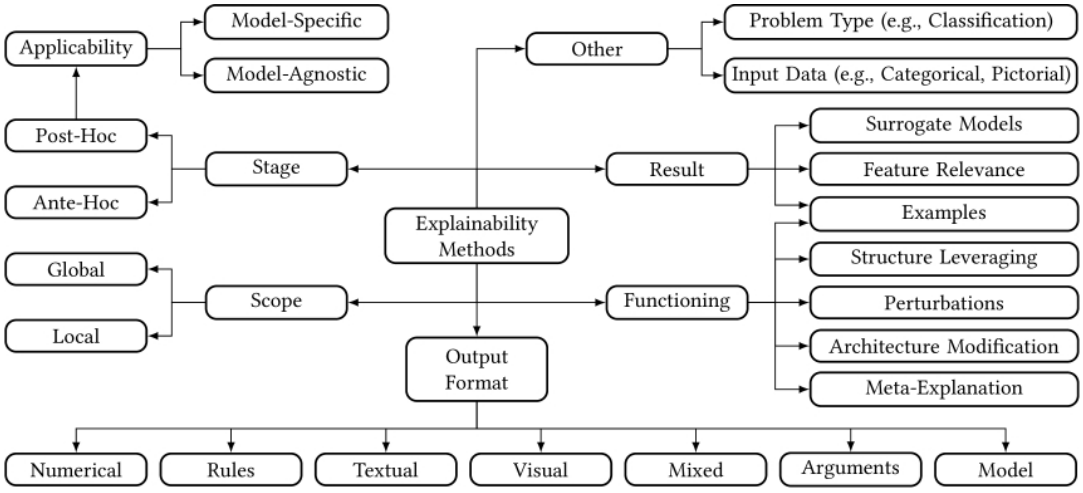


Figure 2: From the work of Speith [39], this taxonomy was proposed as a result on a review on taxonomies in the field of XAI. We refer to the five dimensions of abstraction in section 4 for further explanation.

3) **Functioning**: Describes the different inner functioning of the methods. *Structure Leveraging* methods, for example, rely on modifying specific structural parts of the model they are trying to explain. Some examples include Gradient-weighted Class Activation Mapping (Grad-CAM) [36].
4) **Output Format**: Describes the format of the result that is produced by the explanation method. Note that this is usually strongly linked to the input format, e.g. saliency maps are typical for pictorial data. It is important to note that XAI methods are not limited to a single output format.
5) **Result** The explanation result. Similar to the output format in the sense that both are related to the resulting explanation from an XAI method, yet this dimension describes the result of the XAI method more abstractly. Surrogate modeling, for instance, attempts to find a white-box model (i.e. a model that is inherently interpretable) to approximate a black-box model.

When reviewing the popularity of current research trends, it becomes abundantly clear that current research is shifting towards post-hoc stage algorithms that are model-agnostic [1, 22, 32]. The clear advantage of model-agnostic interpretations is *portability*, or the ability of these interpretations being able to be transported over a larger class of models as opposed to their model-specific counterparts.

5 EXPLAINABILITY DEFINED THROUGH AUDIENCE

To further capture the definition of explainability, previous works typically argue to make a distinction between different *audience groups* [9, 13, 24]. When considering an AI algorithm that classifies medical images, for example, stakeholders would include the patients, doctors, the vendor that created the algorithm and any regulatory agency operating in this domain. Clearly, there is wide variety between these stakeholders in terms of *AI knowledge* and *explanation goal*. The necessity becomes apparent as we note that the technical challenges are strongly tied to this notion. Barredo Arrieta et al. [9] were the first to define explainability with *target audience* as its cornerstone, making a distinction between five groups, with three of these groups appearing in other works as well [24]:

These groups are defined by (1) *End Users*, these are users affected by the decisions of the model, e.g. the patients we mentioned in our example. They are assumed to have little to no knowledge about the data nor the model that made this decision. Their goals are typically an understanding of their personal situation and assessing fairness of the model. (2) *Domain Experts* This group is characterized by users that have knowledge about their specific domain, but not the predictive model they are working with. Their goals include gaining trust in the model, gaining scientific knowledge. (3) *AI Experts*, this group is characterized by knowledge about the model, but not necessarily about the data it is processing. Their explanation need is typically for model assessment and debugging in order to improve the model development process.

This distinction between different users further complicates the concept of explainability. Considering that each user group has different explanation goals and technical knowledge, certain types of explanations will only be appropriate for a specific audience. Trying to argue for a single set of desiderata is likely to be a lost cause; such desiderata would require a modularized scheme. Sokol and Flach [38] propose a variety of dimensions to systematically asses explainable approaches, arguing for a multi-faceted approach to the deployment of explainability methods.

Table 1: Summary of the proposed main challenges of XAI.

Challenge	References	Description
Non-Robust Explanations	[5, 17, 22, 41–43]	Explanations are known to display unexpected behavior in specific settings. As a result, explanations can be unreliable due to unstable or unfaithful explanations.
Hyperparametrizations	[8, 22]	Research on the effect of hyperparameters and their geometric effects on explanations is still lacking. A general understanding on finding optimal hyperparameters would greatly benefit the field.
The Curse of Dimensionality	[22, 41]	High dimensional data can be particularly detrimental to explanations due to the traditional drawback for high dimensionality in ML, but also because of the interactions between ML and XAI.
The Rashomon Effect	[11, 13, 15, 21, 22, 28]	The explanations offered by XAI can be different for models performing equally well on the same dataset. Resulting explanations could be misleading.
Adversarial Attacks	[14, 17–19, 27, 37]	A large body of research has demonstrated that *adversarial attacks* are possible on XAI methods. In this context, adversarial attacks rely on generating perceptually similar instance or models that produce significantly different explanations.
Lack of Quantitative Evaluation	[1–3, 5, 6, 23, 32, 34]	The lack of quantitative evaluation of XAI is another reason why trust is lacking. A properly defined set of metrics to evaluate an explanation of a model should establish a way to assess the performance of XAI algorithms.

6 CHALLENGES OF XAI

In spite of XAI being an appealing concept with a noble goal and vision, the field is still dealing with challenges that are in the way of reaching its full potential. Nevertheless, as discusses in section 2, none of the previous works focus on viewing these challenges purely from the perspective of predictive modeling, and a clear overview is still lacking. Motivated by this concern, our work has focused on identifying challenges in XAI purely related to the field of predictive modeling. As such, we have provided an overview of these challenges in table 1.

7 ROBUSTNESS OF EXPLANATIONS

Misleading interpretations can occur for perturbation-based methods including LIME, Randomized Input Sampling for Explanation of Black-box Models (RISE) Petsiuk et al. [33] and SHAP. These methods work by generating a set of perturbations on instances used for model inference. The predictions generated using these perturbations are in turn aggregated to generate explanations at a local or global level. Because of these randomly generated artificial instances, perturbation-based XAI methods might suffer from issues regarding stability [41–43]. The inherent noise introduced by perturbing instances could lead to this data being non-representative of the local or global space that the perturbed data is trying to emulate. Consequently, produced explanations for statistically similar, or the exact same training data might be different across different iterations for the same XAI method, undermining trust in the explanation as a result.

There are numerous proposed solutions regarding this issue. Zafar and Khan [42] have proposed an alternative approach to Local Interpretable Model-Agnostic Explanations (LIME) by circumventing the data generation step altogether. Instead, the surrogate model is trained solely on the training instances with a weight to each instance based on the distance from the instance to be explained with a Gaussian kernel. Visani et al. [41] propose to instead guarantee stability in regions defined by *stability indices*. They define metrics related to stability and determine regions where these metrics lie within unreasonable values and deem LIME to be unreliable in these regions.

8 HYPERPARAMETRIZATIONS

The effect of hyperparametrizations on XAI explanations has not received a lot of attention in the field. It is clear that high sensitivity of hyperparameters impedes reproducibility, but could also raise question to the correctness of the explanation and ultimately undermine trust [8]. This is especially problematic for the *AI novices* group, because their knowledge about hyperparametrizations is assumed to be negligible. Bansal et al. [8] are one of the few works to have performed sensitivity analysis for various hyperparameters on the effect on the explanations for a set of XAI methods was performed. It was concluded that explanations are relatively sensitive to hyperparameters, i.e. varying the hyperparameters could lead to unstable or unfaithful predictions. Interestingly, it was also shown that XAI methods were surprisingly more robust when explaining a robust model, i.e. a model that is not sensitive to relatively small perturbations in its input.

9 THE CURSE OF DIMENSIONALITY

Considered to be detrimental to machine learning applications in general, XAI is no exception with respect to the curse of dimensionality. We have found that having redundant dimensions for the model can lead to a phletora of issues for a variety of XAI algorithms. This notion becomes even more stringent when the

predictive models can achieve satisfactory prediction scores, even with the redundant dimensions [41].

Getting back to the previous point on *robustness of evaluations*, we first note that the curse of dimensionality aggravates the issue; by increasing the dimensionality, the space of perturbations will become exponentially larger and thus yield more noisy perturbations. As a result, the models fitted on these perturbations will learn this noise instead of the desired local instance and hence yield inconsistent explanations [41]. Computational effort could also suffer drastically in high-dimensional settings. Computing exact *Shapeley values*, for instance, relies on all possible combinations of features, which takes exponential time [31]. Finally, in local methods, definitions of neighbourhood or distance in conjunction with distance metrics could be prone to the curse of dimensionality. Aggarwal et al. [4] show that classic distance metrics such as Euclidean distance scale poorly with dimensionality. As such, using fractured metrics or different distance metrics than Minkowski distances with order of the norm (P value) higher than one was found to yield consistently more effective results.

10 THE RASHOMON EFFECT

The *Rashomon Effect* is the phenomenon that different predictive models with similar performance metrics on the same dataset contradict each other. This is due to the approximation functions being constructed in a different manner. The term was named after the movie "Rashomon" from 1950 and Breiman [11] were the first to formalize the term in the field of predictive modeling. This effect could lead to some contradicting explanations and conclusions about the data. An example is provided by Dong and Rudin [15], who identified a set of equally well performing models for the COMPAS dataset [25]. This dataset concerns itself with predicting recidivism risk and includes sensitive attributes such as race and gender. It was demonstrated that the models differed greatly in the feature importances they attributed. Particularly, they found that the importance of criminal history correlated negatively with the importance of race among different models. Nonetheless, as stated by Hancox-Li [21] "just because race happens to be an unimportant variable in that one explanation does not mean that it is objectively an unimportant variable", i.e. an explanation for a single model might not be a suitable explanation for the actual associations in the data.

We must note that this is not a limitation inherent to XAI; the algorithms perform as they should by staying faithful to predictions of the model. It is expected for different models to have different explanations. Having said that, there are cases in which the Rashomon effect can be used as a means of manipulation, especially considering the *end users* group. Getting back to the COMPAS dataset, it is possible for organisations to opt for models whose explanations place less stress on sensitive factors, such as race, gender and age. Lakkaraju and Bastani [28] conduct a user study in which they demonstrate how such fairwashing practices can be achieved and how they are deemed acceptable by users. Furthermore, in section 11, an *advarsarial attack* will be described which attempts to find a similarly good model that effectively does not explain sensitive features.

The proposed solutions to this effect rely on identifying a set of models that are subject to this effect, which has been conceptualized as the *Rashomon set*. For instance, variable importance clouds could be used to carefully compare and assess the variable importance scores proposed for several models in the Rashomon set [15].

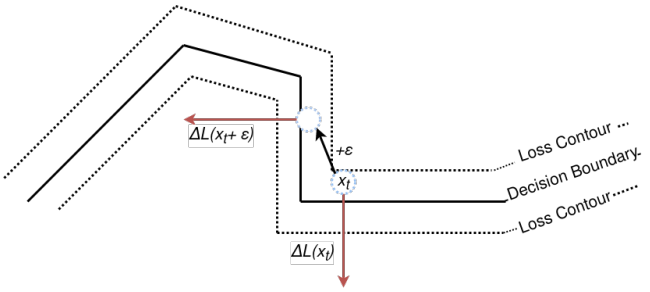
11 ADVERSARIAL ATTACKS

We define *Adversarial attacks* as an umbrella term covering the following two definitions: (1) "Targeted attacks on trained machine learning models using instances that are perceptually indistinguishable, yet produce predictions that are perceptually distinguishable or (2) "Training a model using unrepresentative or inaccurate data in order to generate malicious predictions. For the first of the definitions, the mentioned perturbed instances are usually named *adversarial examples*, while the models mentioned in definition 2 are typically named *adversarial models*. Adversarial attacks were first found out in deep learning (DL) , where classifiers yielded drastically different predictions after being given an adversarial example [18]. In more recent lines of research, adversarial attacks have also been being explored in XAI for DL.

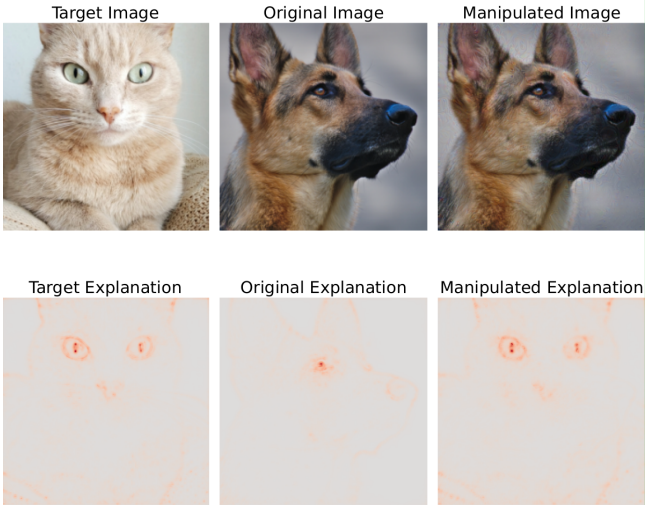
Ghorbani et al. [17] were the first to propose a method to generate such adversarial examples for model-specific methods. The authors argued that the decision boundaries of neural networks with many parameters are roughly piecewise linear [19], and use this intuition to generate their examples. At sharp edges of the decision boundaries, training instances have an especially large influence on the loss while their values are relatively close. After identifying sharp edges in the decision boundaries, training instances near these boundaries are used to generate perceptually infinitesimal perturbations to obtain adversarial examples. A visual example is given in 3a. Whereas earlier works merely demonstrated the existence of adversarial attacks, such attacks yielded no control over the explanation. More recent works have demonstrated that methods exist in order to find adversarial examples that can control explanations arbitrarily [14]. As a proof of concept, such an adversarial example is provided in 3b.

Adversarial machine learning is not limited to adversarial attacks based on instance perturbations that we have demonstrated before. Another proposed idea is building a model which explanations do not convey the actual associations in the data. This attack is based on the *Rashomon effect*, which we have elaborated upon in section 10. Such models are named *adversarial models*, where instead of the training instances, the model is manipulated to generate biased explanations. Slack et al. [37] have introduced such an adversarial model in order to attack LIME and SHAP. An adversarial model was created on the COMPAS dataset [25], which has the goal of predicting recidivism and harbors sensitive features such as race, gender and age. LIME and SHAP generate perturbations that do not consider the distribution between features in the dataset which causes these perturbations to be out-of-distribution with respect to the original data. Consequently, instances that are deemed out-of-distribution could be trained using a classifier trained on innocuous features with zero correlation to discriminatory features, while the other instances could be trained with a (racially) biased model that only considered such discriminatory features. Given that the real-world data was assumed to follow the same distribution that

the racially biased model was trained on, the resulting model was deemed to be racially biased. Nonetheless, the explanations generated for this model were not showing the racially biased features as important to the model, in spite of the model being biased by design.



(a) The intuition behind adversarial attacks.



(b) An example of an adversarial attack.

Figure 3: Inspired by the work of Ghorbani et al. [17], figure 3a presents the intuition behind adversarial attacks. At areas where the decision boundaries are non-smooth, training points have a large effect on the loss gradient. Consider instance x_t perturbed by ϵ, it can be observed that this new instance $x_t + \epsilon$ has a gradient perpendicular to its counterpart and thus a signifcantly different explanation. Figure 3b is a concrete example of an adversarial attack. Note that the original image is similar to the manipulated one, yet the explanation is manipulated.

12 LACK OF QUANTITATIVE EVALUATION

Properly quantifying the behaviour of model explanations is a challenging task. Firstly, there is no such thing as ground truth in this field; if there were to be a ground truth, there would no need for explanations. Secondly, it is unclear which of the phletora of methods to select. Motivated by these challenges, Nauta et al. [32] have compiled a systematic literature review on evaluating explainable AI. Some of their main findings included that many of the recent works still either rely solely on user studies and anecdotal evidence to report their results. Specifically, in only 58% of the papers that were analysed, some form of quantitative evaluation was applied. Nevertheless, it was concluded that the amount of studies using evaluation metrics have been steadily increasing, in comparison to the previous studies [1]. Using anecdotal evidence and human evaluation instead of objective, quantitative evaluation is deemed to be misleading in assessing explainability methods is furthermore deemed to be misleading according to several papers. Petsiuk et al. [33] argue that "keeping humans out of the loop for evaluation makes it more fair and true to the classifier's own view on the problem rather than representing a human's view" and Adebayo et al. [2] have found that some explanations can be independent of the model that trying to explain without humans being aware. Similarly, [29] argue that interpretability research as a whole "suffers from an over-reliance on intuition-based approaches that risk –and in some cases have caused– illusory progress and misleading conclusions."

12.1 Evaluation Schemes

In line with the maturation of the XAI field, several evaluation schemes have been suggested by the research community in more recent research works. We have identified a few of these methods below:

1) **Co-12 categorization scheme** [32]: Besides providing a systematic literature review, Nauta et al. [32] also provided a categorization scheme in order to define evaluation qualitatively. Explainability is described as being multi-faceted and this is made explicit through a set of 12 properties, which are called *co-12 properties*. For each of these properties, relevant quantitative methods are classified to their appurtenant co-12 property. Through this comprehensive overview, the authors aimed to provide a more inclusive view of explainability in order to use quantitative evaluation in an insightful manner.
2) **Faithfulness** [6]: The concept of *faithfulness* is unavoidable in the discussion of evaluation metrics. This describes the idea of how closely an explanation follows the underlying model. Alvarez-Melis and Jaakkola [6] have defined the metric of faithfulness in order to evaluate the Pearson correlation between the importance scores to the actual attribution of features towards the prediction. In order to achieve this, features deemed important by the XAI method are incrementally removed and predictions are made with the reduced set of features. Petsiuk et al. [33] propose similar methods for image data, adding on by describing a method of continuously adding features to measure the effect on predictive accuracy.
3) **RemOve And Retrain ROAR** [23] is a proposed evaluation benchmark specifically aimed towards model-specific feature relevance based approaches. Similar to *Faithfulness*, this method relies on removing features and evaluating the

act on the performance of the model. The key difference is that ROAR removes and fully retrains the model with these features. It was found that a majority of explanation methods did not perform better than or was on par with a random assignment of feature importances. Only *VarGrad* and *SmoothGrad-Squared* were able to outperform this random baseline.

4) **OpenXAI** [3] is an open source framework for evaluating and benchmarking post-hoc explanation methods. real-world and synthetic datasets and a set of 22 evaluation metrics were used in order to achieve this. It was argued that this method is novel in the sense that it was the first to encapsulate three notions of explanation reliability that were identified in previous works: *faithfulness*, *stability* and *fairness*. Additionally, benchmarks across eight different datasets have been performed in order to compare the explanation reliability of six different post-hoc XAI methods.
5) **L**ocal **E**xplanation ev**A**luation **F**ramework (LEAF) [7] is a proposed set of metrics to compare and evaluate models based on *feature importance* (see Figure 2 for a taxonomic overview). Their work focusesd specifically on evaluating two of the most-used model-agnostic algorithms, namely LIME and SHAP. Their metrics included the two more commonly used *conciseness* and *local fidelity* and the novel *local concordance*, *reiteration similarity* and *prescriptivity*. A detailed experimental evaluation on a variety of datasets was performed, comparing LIME with SHAP.

13 DISCUSSION

From our review we were able to compile a body of works that were largely published at prominent venues of the XAI community, including FAccT, NeurIPS, SIGKDD, AAI and CVPR. A small minority of our work is composed of pre-prints that carried relevance in the field were analysed and included. Admittedly, we intentionally reduced the volume of research from our initial searches to allow for proper analysis of the selected papers, i.e. sacrificing breadth in order to gain depth in our analysis. As a result, there are no quantitative guarantees that this work encompasses all the main challenges, hence we explicitly do not claim our list of challenges to be exhaustive.

Another thing worth noting is that the provided challenges and examples are not mutually exclusive. When considering the Rashomon effect and the adversarial attack described against model-agnostic perturbation-based methods, we note that such an attack may also be defined as finding a model in the Rashomon set. Furthermore, we notice the intersection between non-robust explanations, hyperparametrisations and the curse of dimensionality. In essence, the idea connecting these challenges is that the noise introduced by generating random perturbations for the algorithm could transfer some of this noise into the final explanation. We have noted that hyperparameters and dimensionality also play a role in the stability of explanations.

Combing back to the research questions, as defined in 3.2, we have identified a body of main challenges in 1 through a systematic search of literature within XAI that addresses *RQ1*. Each of these questions was presented in a more in-depth manner consecutively. Furthermore, *RQ2* was defined as "Are there studies suggesting that XAI methods have predictable behaviour?". Through our identified works, it was concluded that the opposite of this ; most works concerning themselves with studying the behaviour of XAI have found that XAI behaves unpredictable in specific cases, as has been demonstrated through *adversarial attacks*, *high dimensional*, *hyperparameterizations*. Finally, our third question: "Is there a methodology for ensuring the reliability of XAI methods?" is essentially answered through providing theoretical guarantees using quantitative evaluation. However, the field on XAI has not agreed on unified metrics for evaluation hitherto, which is impeding credibility of explanations. In subsection 12.1, we have provided some of the recent works that provide such evaluation schemes. We can further conclude RQ3 by noting that the field of evaluation for XAI is maturing as evaluation metrics are increasingly adopted in research.

14 CONCLUSION & FUTURE WORK

This survey revolved around outlining the field of eXplainable Artificial Intelligence (XAI), which is deemed to be paramount step towards the adoption of ML models in mission-critical domains. In order to do so, a body of research was identified. This work has focused on providing a high-level overview of the algorithmic landscape and briefly touching upon the concept of model explainability and its strong ties to audience. Our analysis has yielded a global overview of the core technical challenges that the field is currently facing identified through literature that addresses these limitations. A more in-depth description of each challenge was subsequently provided, describing the intuition behind the challenges, indicating its relevance through proof of concepts or examples on their potential industry disruption and indicating some proposed solutions.

Through identifying challenges in XAI, we hope to outline some opportunities for future work. We concretely advocate for directly addressing the core challenges in 1. Some more specific directions include providing further experimental evaluation on the effect of hyperparameters on explanations, i.e. providing theoretical analyses on hyperparametrizations or optimizing in more general settings, (2) Making perturbation-based methods more resistant against adversarial attacks through more intelligent sampling, (3) devising methods to detect adversarial attacks for explanations, (4) unifying evaluation metrics into a single benchmark or (5) Devising new methods to help mitigate the Rashomon effect through finding models in the Rashomon set and interpreting these results.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

I would like to thank my supervisor dr. Elena Ranguelova for supervising this work, providing relevant insights and also being the one to see the rationale for and incentivise this work. I would also like to thank dr. Sonja Georgievska for providing additional insights and discussion on the matter.

REFERENCES

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

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 9525–9536.

[3] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. 2022. OpenXAI: Towards a Transparent Evaluation of Model Explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=MU2495w47rz

[4] Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory (ICDT '01)*. Springer-Verlag, Berlin, Heidelberg, 420–434.

[5] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. https://doi.org/10.48550/ARXIV.1806.08049

[6] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards Robust Interpretability with Self-Explaining Neural Networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 7786–7795.

[7] Elvio Gilberto Amparore, Alan Perotti, and Paolo Bajardi. 2021. To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science* 7 (2021).

[8] Naman Bansal, Chirag Agarwal, and Anh Nguyen. 2020. SAM: The Sensitivity of Attribution Methods to Hyperparameters. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8670–8680. https://doi.org/10.1109/CVPR42600.2020.00870

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[11] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (2001), 199 – 231. https://doi.org/10.1214/ss/1009213726

[12] Arun Das and Paul Rad. 2020. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. *ArXiv* abs/2006.11371 (2020).

[13] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* 39, 2 (2022), 101666. https://doi.org/10.1016/j.giq.2021.101666

[14] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4a-Paper.pdf

[15] Jiayun Dong and Cynthia Rudin. 2020. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence* 2 (12 2020), 810–824. https://doi.org/10.1038/s42256-020-00264-0

[16] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1 (2018), eaao5580. https://doi.org/10.1126/sciadv.aao5580 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580

[17] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of Neural Networks is Fragile. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 452, 8 pages. https://doi.org/10.1609/aaai.v33i01.33013681

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. https://doi.org/10.48550/ARXIV.1412.6572

[19] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6572

[20] Hajar Hakkoum, Ibtissam Abnane, and Ali Idri. 2022. Interpretability in the medical field: A systematic mapping and review study. *Applied soft computing* 117 (2022), 108391–.

[21] Leif Hancox-Li. 2020. Robustness in Machine Learning Explanations: Does It Matter?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 640–647. https://doi.org/10.1145/3351095.3372836

[22] Andreas Holzinger, Randy Goebel, Ruth Fong, Taesup Moon, Klaus-Robert Müller, and Wojciech Samek. 2022. *xxAI - Beyond Explainable Artificial Intelligence*. 3–10. https://doi.org/10.1007/978-3-031-04083-2_1

[23] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. *A Benchmark for Interpretability Methods in Deep Neural Networks*. Curran Associates Inc., Red Hook, NY, USA.

[24] Zhongli Filippo Hu, Tsvi Kuflik, Ionela Georgiana Mocanu, Shabanam Najafian, and Avital Shulner Tal. 2021. Recent Studies of XAI - Review. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 421–431. https://doi.org/10.1145/3450614.3463354

[25] Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm. https://web.archive.org/web/20230406011138/https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm Accessed: 06-04-2023.

[26] T. Jones. 2023. *Pause Giant AI Experiments: An Open Letter.* https://web.archive.org/web/20230330110247/https://futureoflife.org/open-letter/pause-giant-ai-experiments/ Accessed: 2023-03-30.

[27] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2022. *The (Un)Reliability of Saliency Methods*. Springer-Verlag, Berlin, Heidelberg, 267–280. https://doi.org/10.1007/978-3-030-28954-6_14

[28] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[29] Matthew L. Leavitt and Ari Morcos. 2020. Towards falsifiable interpretability research. https://doi.org/10.48550/ARXIV.2010.12016

[30] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[31] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

[32] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* (feb 2023). https://doi.org/10.1145/3583558 Just Accepted.

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[34] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[35] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[36] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74

[37] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830

[38] Kacper Sokol and Peter Flach. 2020. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 56–67.

https://doi.org/10.1145/3351095.3372870

[39] Timo Speith. 2022. A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2239–2250. https://doi.org/10.1145/3531146.3534639

[40] Giulia Vilone and Luca Longo. 2021. Classification of Explainable Artificial Intelligence Methods through Their Output Formats. *Machine Learning and Knowledge Extraction* 3, 3 (2021), 615–661. https://doi.org/10.3390/make3030032

[41] Giorgio Visani, Enrico Bagli, Federico Chesani, Alessandro Poluzzi, and Davide Capuzzo. 2022. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models. *Journal of the Operational Research Society* 73, 1 (2022), 91–101. https://doi.org/10.1080/01605682.2020.1865846 arXiv:https://doi.org/10.1080/01605682.2020.1865846

[42] Muhammad Rehman Zafar and Naimul Khan. 2021. Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability. *Machine Learning and Knowledge Extraction* 3, 3 (2021), 525–541. https://doi.org/10.3390/make3030027

[43] Zhengze Zhou, Giles Hooker, and Fei Wang. 2021. S-LIME: Stabilized-LIME for Model Explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2429–2438. https://doi.org/10.1145/3447548.3467274