Machine Learning post-hoc interpretability: a systematic mapping study

Carla Piazzon Ramos Vieira
carlaprv@usp.br
University of São Paulo
São Paulo, São Paulo, Brazil

Luciano Antonio Digiampietri
digiampietri@usp.br
University of São Paulo
São Paulo, São Paulo, Brazil

ABSTRACT

Context: In the pre-algorithm world, humans and organizations made decisions in hiring and criminal sentencing. Nowadays, some of these decisions are entirely made or influenced by Machine Learning algorithms. **Problem:** Research is starting to reveal some troubling examples in which the reality of algorithmic decision-making runs the risk of replicating and even amplifying human biases. Along with that, most algorithmic decision systems are opaque and not interpretable - which makes it more difficult to detect potential biases and mitigate them. **Solution:** This paper reports an overview of the current literature on machine learning interpretability. **IS Theory:** This work was conceived under the aegis of the Sociotechnical theory. Artificial Intelligence systems can only be understood and improved if both 'social' and 'technical' aspects are brought together and treated as interdependent parts of a complex system. **Method:** The overview presented in this article has resulted from a systematic mapping study. **Summary of Results:** We find that, currently, the majority of XAI studies are not for end-users affected by the model but rather for data scientists who use explainability as a debugging tool. There is thus a gap in the quality assessment and deployment of interpretable methods. **Contributions and Impact in the IS area:** The main contribution of the paper is to serve as the motivating background for a series of challenges faced by XAI, such as combining different interpretable methods, evaluating interpretability, and building human-centered methods. We end by discussing concerns raised regarding explainability and presenting a series of questions that can serve as an agenda for future research in the field.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → **Machine learning approaches**.

KEYWORDS

xai, machine learning, explainability, interpretability, fairness, black-box

SBSI 2022, June 7–10, 2021, Curitiba, Brazil

ACM ISBN 978-1-4503-9698-1/22/05…$15.00
https://doi.org/10.1145/3535511.3535512

ACM Reference Format:
Carla Piazzon Ramos Vieira and Luciano Antonio Digiampietri. 2022. Machine Learning post-hoc interpretability: a systematic mapping study. In *XVIII Brazilian Symposium on Information Systems (SBSI 2022), May 16–19, 2022, Curitiba, Brazil.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3535511.3535512

1 INTRODUCTION

Machine Learning algorithms have penetrated every aspect of our daily lives. Algorithms make movie recommendations, suggest products to buy, and are increasingly used in high-stakes decisions in loan applications, dating, and hiring. However, Gebru [10] says that this rapid permeation of Artificial Intelligence (AI) into society has not been accompanied by a thorough investigation of the sociopolitical issues that cause certain groups of people to be harmed rather than advantaged by it. Books such as Weapons of Math Destruction [26] details how people in lower socioeconomic classes in the US are subjected to more automated decision-making tools than those who are in the upper class.

While the very first AI systems were easily interpretable, the last years have witnessed the rise of opaque decision systems such as Deep Neural Networks (DNNs). As humans and companies rely more and more on complex artificial intelligence models, providing explanations for their decisions to support an effective human-computer interaction becomes increasingly important.

In order to avoid society being harmed or marginalized by the current generation of AI systems, interpretability proposes creating a suite of Machine Learning techniques that can 1) help ensure fairness in decision-making by detecting biases, 2) produce more explainable models while maintaining performance, and 3) enable humans to understand and appropriately trust artificial-intelligence-based systems.

For these reasons, research on eXplainable Artificial Intelligence (XAI) [14] gained significant relevance. The number of scientific papers and conferences around the world about Machine Learning Interpretability has significantly increased over the last decade. This direction is confirmed by the establishment of projects such as DARPA's Explainable AI [14], the European response to the General Data Protection Regulation [12], and the recent letters from IBM, Amazon, and Microsoft that decided to stop research and development of facial recognition technologies.

Regarding Interpretability methods an important distinction can be made: self-explaining and post-hoc approaches. Self-explaining approaches refer to developing machine learning models that are considered inherently interpretable such as decision trees, linear models, and so on. In contrast, the post-hoc approach requires creating a second model to provide explanations for a black-box

model. In this article, we provide an overview of the current literature on machine learning post-hoc interpretability by conducting a systematic mapping study.

This study seeks to provide newcomers to the field of interpretability an overview that can serve as reference material and help find a suitable model to solve interpretability problems more easily by (i) categorizing existing methods, (ii) presenting limitations of current interpretability methods, and (iii) discussing challenges that are still to be addressed in the field. In the Related Work section, we present the related works. In the Methodology section, we describe the research methodology used in our study. In the Results section, we present the interpretability techniques and evaluations metrics found during this study. In the Discussion and challenges for future research, we discuss challenges still to be addressed in the field and list future research directions. Finally, in the Conclusions section, we present our conclusions.

2 RELATED WORK

During our search for related works, we found out that explainability is not a new area of study as it may seem. Single-tree approximations for Neural Networks were first presented in 1995 by Craven and Shavlik [7].

The survey presented by Guidotti et al. [13] outlined a taxonomy to provide classifications of the main problems with opaque algorithms, examining four features for each explanation method. Other works include clarifying concepts related to interpretability [4], distinguish between interpretability and explainability [11], and proposing goals for explainability [3].

Furthermore, Carvalho et al. [6] discussed the impact of interpretable methods on social development and Du et al. [8] divided the current methods into two groups: "global" and "local". While global understanding is particularly important for assessing trust in a model as a whole (before deployment), most of the current research has been devoted to explaining individual predictions.

The classification of existing work for interpretable methods mainly focuses on two categories: intrinsic interpretability and post-hoc interpretability [6, 8].

Barredo Arrieta et al. [2] provide an extensive review on explainable AI, where concepts and taxonomies are presented, and challenges are identified. While their review covers Machine Learning interpretability in general, our review is specific to post-hoc methods and offers unique perspectives and insights.

Specifically, our review provides new perspectives in the following senses: 1) We treat post-hoc and intrinsic interpretability separately because the former explains an existing black-box model, while the latter constructs interpretable models; 2) we propose a new classification of interpretability methods considering its explanation technique and 3) we provide important sociotechnical considerations in order to build human-centered solutions. Interpretability research is a rapidly evolving field, and many research articles are produced every month. Therefore, our review should be a valuable addition to the literature.

3 METHODOLOGY

The results reported in this paper are the result of a systematic mapping study. Systematic Mapping Study (SMS) is used to structure a research area, identify gaps that are not covered by the current researches, and offers insights for future work in the area. The main parts of the protocol that guided the study are presented in the following.

3.1 Research questions

The research questions were used to guide the research. The following research questions (RQ) were defined:

- **1**: Which strategies and methods have been applied to interpret machine learning models?
- **2**: Which metrics have been used to evaluate interpretability methods?
- **3**: Which are the challenges still to be addressed in the field?

3.2 Search strategy

3.2.1 Data sources and search string. The search for the articles was conducted using three digital libraries: IEEExplore, ACM Digital Library, and Scopus. The following search expression was applied in title, abstract, and keyword fields.

Table 1: Search String

(("black-box" OR "black box") AND ("machine learning" OR "predict*" OR "model*" OR "classif*") AND ("explainable" OR "explainability" OR "interpretability") AND ("agnostic" OR "post-hoc" OR "post hoc"))

We have used a group of known articles manually collected in order to define the words of the search string. The articles are listed below:

- A Study on Interpretability of Decision of Machine Learning [32]
- Survey of Methods for Explaining Black Box Models [13]
- Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods [34]
- What's inside the black-box? [9]
- "Why Should I Trust You?": Explaining the Predictions of Any Classifier [29]

3.2.2 Inclusion and Exclusion Criteria. Each study returned in the search phase was analyzed by reading its title and abstract. Through this analysis, duplicated studies were identified. Moreover, studies that didn't match at least one of the following inclusion criteria were rejected:

- **IC-1 Peer-reviewed paper**: the paper must have been published in journals or conferences with peer review.
- **IC-2 Primary study**: only papers providing a direct contribution on XAI (e.g., models, architectures, or implementations) are included, secondary studies (i.e. surveys) are excluded.
- **IC-3 Not restricted access**: the content of the article should be accessible and should not have restricted access.
- **IC-4 Written in English**: the paper must have been published in English.

- **IC-5 Related to Artificial Intelligence**: the paper must be related to Artificial Intelligence and co-related areas.
- **IC-6 Explanations relevant to AI field**: papers addressing explanations in other areas without any relevancy to AI are excluded.
- **IC-7 Tabular and labeled data**: the paper proposes a method for tabular and labeled data.
- **IC-8 Post-hoc interpretability method**: the paper proposes a model-agnostic approach.

The inclusion criteria defined intend to assure some important aspects of the collected studies: quality *IC-1*, primarity *IC-2*, accesibility (*IC-3* and *IC-4*) and scope (*IC-5*, *IC-7* and *IC-8*).

We have selected only papers that deal with tabular and labeled data as the interpretability approaches for images/video or text inputs would wider the scope of this study. Also, the criteria *IC-8* is important as our focus is not reviewing specific techniques (e.g: explaining neural networks only) but techniques that deal with different types of models.

3.2.3 Quality score. Due to the scope of this study, we have defined a quality score comprised of four categories to evaluate the level of reproducibility of the selected studies. Each category corresponds to a reproducibility aspect of research design and is associated with a numerical score, as follows:

- Methodology: indicates if the study methodology was clearly explained (1 - yes, 0 - no)
- Code: is the source code of the study available? (1 - code or pseudo-code, 0.5 - algorithm, 0 - not available)
- Data: is the data used by the authors available for reuse? (1 - yes, 0 - no)
- Examples: indicates if explanations generated by the interpretability method proposed are present in the article (1 - yes, 0 - no)

The selected studies were sorted in descending order of quality score and only the ones with the highest scores were fully read by the authors. Studies that received a total score of 3.5 (or higher) were selected for this study.

3.3 Data extraction and data synthesis

The data extraction phase was performed with papers accepted in the selection phase. All studies selected were fully read in order to extract the information using the form shown in Table 2.

3.4 Conduction results

This study reports the results based on searches performed in June 2020. A total of 163 papers were found using the search string in the three digital libraries. During the selection phase, 29 duplicated studies were identified, 121 papers were rejected according to the inclusion criteria, then, 2 papers were rejected based on the quality score. Table 3 presents the distribution of the excluded studies according to the inclusion criteria.

Most of the rejected papers did not match the inclusion criteria (*IC-6: papers addressing explanations in other areas without any relevancy to AI are excluded*). We identified that these studies were about cybersecurity, chatbots, and software testing. Furthermore, among the excluded studies, we found studies that deal with multimedia data (images, videos, or audio) which is not according to criteria (*IC-7: the papers proposes a method for tabular and labeled data*).

Also, the studies had their quality score calculated. Only the papers with a quality score equal to or higher than 3.5 were considered, resulting in 11 publications. The quality scores of studies included are presented in Table 4.

After the selection phase, 11 studies were accepted for the data extraction phase. Despite presenting a relatively high rejection, we considered the search string as valid because it enabled us to identify studies that propose methods pertinent and inside the scope of this study.

In the data extraction phase, articles were fully read and information about each study was extracted.

Table 2: Data extraction form

Publication information
- Title
- Authors
- DOI/Link
- Keywords
- Year of publication
Methods
- Study objective as stated by the authors
- Interpretability method name
- Black-box models used for evaluation
- Scope of the interpretability method (local and/or global)
- Which strategies were used to extract explanations?
- Explanation format
Datasets
- Dataset name and brief description
- Dataset source
Evaluation
- Which metrics were used to evaluate the method?
- Metric name and formal definition
- Metric values and benchmark used
- User-studies conducted
Future research directions

Table 3: Inclusion criteria application result

Criteria	Description	Excluded studies (%)
IC-1	Peer-reviewed paper	0.0%
IC-2	Primary study	20.7%
IC-3	Not restricted access	2.5%
IC-4	Written in English	0.0%
IC-5	Related to Artificial Intelligence	0.0%
IC-6	Explanations relevant to AI field	47.9%
IC-7	Tabular and labeled data	20.7%
IC-8	Post-hoc interpretability method	8.3%

Table 4: Quality scores of studies included in the systematic review

Tool	Reference	M	C	D	E	Score
G-REX	[16]	1	0.5	1	1	3.5
MUSE	[18]	1	0.5	1	1	3.5
GP	[9]	1	0.5	1	1	3.5
MAPPLE	[27]	1	1	1	1	4
SHAP	[22]	1	0.5	1	1	3.5
LIME	[29]	1	1	1	1	4
GAM	[15]	1	1	1	1	4
Fair-MAML	[33]	1	1	1	1	4
MC-BRP	[21]	1	1	1	1	4
CERTIFAI	[31]	1	0.5	1	1	3.5
DiCE	[25]	1	1	1	1	4

M - Methodology, C - code, D - Data, E - Examples

4 RESULTS

This section presents the collected results regarding the formulated research questions.

4.1 RQ1: Which strategies and methods have been applied to interpret machine learning models?

Explanation methods and techniques for Machine Learning interpretability can be classified according to different criteria. In this study, we propose the criteria described, as follows:

(1) **Stage:** it refers to the stage at which a method generates explanations: intrinsic or post-hoc.
(2) **Agnosticity:** interpretability method can be either model-agnostic or model-specific.
(3) **Scope:** it refers to the scope of an explanation that can be either global or local.
(4) **Explanator**: it defines the strategy used to extract explanations from a black-box model.

The advantage of the proposed division is that it highlights the characteristics of different approaches and helps find the most suitable interpretability method to a machine learning problem.

Table 5 presents a summary of the interpretability methods considering the defined criteria.

4.1.1 ***Stage: Intrinsic x post-hoc.*** This criterion distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyze the model after its training (post-hoc) [24]. Intrinsic interpretability refers to machine learning models that are considered interpretable due to their simple structure. The family of this category includes decision trees, rule-based models, linear models, and so on. In contrast, post-hoc interpretability requires creating a second model to provide explanations for an existing model.

4.1.2 ***Agnosticity: Model agnostic x Model specific.*** Model-specific interpretation methods are limited to specific model classes. The interpretation of regression weights in a linear model is a model-specific interpretation since the interpretation of intrinsically interpretable models is always model-specific. Methods that only work for the interpretation of e.g. neural networks are model-specific. Model-agnostic methods can be used on any machine learning model and are applied after the model has been trained (post-hoc). By definition, these methods cannot have access to model internals such as weights or structural information.

4.1.3 ***Scope: Local x Global.*** Based on previous categorizations, we can further differentiate two types of interpretability: global and local interpretability. In the former case, the goal is to make the entire inferential process of a model transparent and comprehensible as a whole. In the latter case, the objective is to explicitly explain each inference of a model (or any specific inference of the model).

4.1.4 ***Explanator.*** Model-agnostic techniques for post-hoc explainability are designed to be plugged into any model in order to extract some information from its prediction process. Different techniques can be used to achieve that and they rely on model simplification, feature relevance, and counterfactual examples:

Explanations by simplification. This group of techniques is the broadest under the category of model agnostic post-hoc methods. It refers to the techniques that approximate an opaque model using a simpler one (surrogate). The main challenge comes from the fact that the simple model has to be flexible enough so it can approximate the complex model accurately.

Local explanations are present in this category. For example, Local Interpretable Model-Agnostic Explanations (LIME) [29] explains the prediction of any classifiers by learning a local self-interpretable model (such as linear models or decision trees), trained on a new dataset that contains interpretable representations of the original data.

However, almost all techniques in this group produce rule or decision-set-based explanations by exploiting several rule-extraction techniques.

The method Genetic Rule EXtraction (G-REX) [16] employed genetic algorithms to generate IF-THEN rules with AND/OR operators. This algorithm starts from an empty set of rules and adds, at each iteration, a rule for each feature predicate. This method identifies the candidate rules with the highest estimated precision over a dataset where precision is equal to the proportion of correct predictions.

MUSE [18] is a method that creates sets of IF-THEN rules. MUSE is based on an objective function that simultaneously optimizes accuracy and interpretability by learning short and compact decision sets that capture the behavior of a given black-box model, cover the whole feature space, and pay attention to small but important classes.

MAPLE (Model Agnostic SuPervised Local Explanations) [27] combines the idea of using random forests as a method for supervised neighborhood selection for local linear modeling, introduced by Bloniarz et al. [5] as SILO, with the feature selection method proposed by Kazemitabar et al. [17] as DStump. This method generates local explanations provided by the supervised neighborhood approach and global explanations by rule extraction.

Evans et al. [9] propose a global model extraction that uses Genetic Programming as a tree construction method: rather than trees being constructed in a top-down manner, trees are evolved from a population of candidates.

Table 5: Interpretability Methods

Tool	Reference	Stage	Scope	Agnosticity	Explanator
G-REX	[16]	Post-hoc	Local	Agnostic	Explanation by simplification
MUSE	[18]	Post-hoc	Global	Agnostic	Explanation by simplification
GP	[9]	Post-hoc	Global	Agnostic	Explanation by simplification
MAPPLE	[27]	Post-hoc	Local	Agnostic	Explanation by simplification
LIME	[29]	Post-hoc	Local	Agnostic	Explanation by simplification
SHAP	[22]	Post-hoc	Local	Agnostic	Feature importance explanation
GAM	[15]	Post-hoc	Global	Agnostic	Feature importance explanation
Fair-MAML	[33]	Post-hoc	Local	Agnostic	Feature importance explanation
MC-BRP	[21]	Post-hoc	Local	Agnostic	Explanation by example
CERTIFAI	[31]	Post-hoc	Local	Agnostic	Explanation by example
DiCE	[25]	Post-hoc	Local	Agnostic	Explanation by example

Feature relevance explanations techniques aims to clarify the inner functioning of a model by computing a relevance score for its input variables. These scores quantify the sensitivity a feature has upon the output of the model. The assumption is that the larger the difference in the outcome, the more relevant the feature is for the model's prediction.

One commonly used technique within this group is Shapley Additive exPlanations (SHAP) [22]. SHAP is a game-theoretic approach to explain individual predictions by computing the contribution of each feature to the prediction. SHAP values measure the impact of having a certain value for a given feature in comparison to the prediction we'd make if that feature took some baseline value.

Ibrahim et al. [15] propose a method called Global Attribution Method (GAM) that's able to explain nonlinear representations learned by neural networks across subpopulations. The authors argue that existing global techniques rely on more interpretable surrogate models such as decision trees or manipulate the input space to assess global predictive power. This approach address one of LIME's drawbacks related to the non-linearity of complex models.

GAM groups similar local feature importance to form human-interpretable global attributions. Each local attribution is treated as a ranking of features and the elements of this attribution vector correspond to the importance of the feature for a particular prediction. The attributions are transformed into normalized percentages to consider only feature importance, not whether a feature contributes positively or negatively to a particular target.

Fairness Warnings [33] provides an interpretable model that predicts which changes to the testing distribution will cause a model to behave unfairly (mean shifts in test distribution). Fair-MAML is a meta-learning method that "learns to learn" fairly and can be used to train a fair model quickly from minimal data.

The approach uses Supersparse Linear Integer Models (SLIM) as the interpretable model. SLIM creates a linear perceptron that reduces the magnitudes of the coefficients, removes unnecessary coefficients, and forces the coefficients to be integers.

Compared to those attempting explanations by simplification, a similar amount of publications were found regarding explainability by means of feature relevance techniques. Implying that as well as with model simplification techniques, feature relevance has also become a popular subject study in the XAI field.

Explanations by example: Counterfactual explanations were first introduced by Wachter et al. [35] as a way of explaining model results to users such that they can understand why a particular decision was made and identify what would need to change in order to receive the desired result in the future. Given an input instance and a classifier model, a counterfactual explanation is defined as a generated instance that is close to the input instance but for which the model gives a different output. It is used to describe a causal situation in the form: "If A had not occurred, B would not have occurred".

Sharma et al. [31] introduce Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models (CERTIFAI). Most of the methods that generate counterfactual explanations are limited to linear models. CERTIFAI generates counterfactual explanations via a custom genetic algorithm. The meta-heuristic evolutionary algorithm starts by generating a random set of data points that have different predictions from the input data point. A subsequent evolutionary process results in a set of data points close to the input that maintain the prediction constraint.

However, it is difficult to generate counterfactual examples that are actionable for a user situation. Using a loan decision example, a CF explanation may suggest to "change your house rent", but it does not say much about alternative counterfactual, or consider the relative ease between different changes a user may need to make. CERTIFAI provides the user some kinds of constraints that help make the counterfactual more realistic:

(1) *Muting features:* define features that cannot have their values changed;
(2) *Feature range:* define the range of specific features. For example, it might be difficult for a user to drastically increase their income to be approved for a loan, so an income range can be specified;
(3) *Number of explanations:* users can specify how many such explanations the method should generate.

These auxiliary constraints are incorporated by restricting the space from which individuals can be generated, to ensure feasible solutions. Ideally, these examples should balance between a wide range of suggested changes (diversity), and the relative ease of

adopting those changes, and also follow the context of society, e.g., one can hardly lower their educational degree or change their race.

Mothilal et al. [25] propose a method (DiCE) that generates sets of diverse counterfactual examples for any machine learning classifier. Such as CERTIFAI, DiCE supports generating a set of counterfactual explanations and simple constraints on features to ensure the feasibility of the generated counterfactual examples.

Monte Carlo Bounds for Reasonable Predictions (MC-BRP) [21] differs from the others described in this section as it does not aim to identify one counterfactual example, but rather a range of feature values for which the prediction would be different.

4.2 RQ2: Which metrics have been used to evaluate interpretability methods?

Although initially considered for rule extraction methods [7], we might consider the following dimensions for evaluating models in terms of explainability:

- **Fidelity [22, 28]:** it refers to how well does the explanations generated approximate the prediction of the black-box model. It is measured in terms of accuracy, F1-score, and so on, but concerning the outcome of the black-box. Given a data set D = X we can apply to each record x ∈ X both the model: (i) for the black-box b we get the set of predictions Y, and (ii) for the interpretable predictor i we get the set of predictions Z. Thus, fidelity score can be calculated by applying the same calculus of the accuracy function where the target values are the predictions Y of the black-box b against the predicted values Z. Accuracy and fidelity are closely related. If the black-box model has high accuracy and the explanations generated have high fidelity, these explanations also have high accuracy. Messalas et al. [22] argue that high fidelity does not necessarily imply that the decision process of the two models is the same. By decision process, they mean the features that the model relied on, in order to make a prediction. So, they have introduced a new metric, "Top Similarity", which measures the internal fidelity between the original and the surrogate model. "Top Similarity" is not only measured in terms of accuracy, but it analyses the similarity between variables importance of each model.
- **Comprehensibility:** The extent to which extracted representations are humanly comprehensible. Some studies explored ways in which the end-users can contribute to minimizing misclassifications. Ribeiro et al. [28] simulated user trust in explanations by defining "untrustworthy" explanations and models. A different approach in quantifying explanations quality with human intuition has been taken by Lakkaraju et al. [18] that defined an explanation quality metric based on user task completion time and accuracy of the answers provided by users. Another example is the work of Messalas et al. [22], who compared SHAP, LIME, and DeepLIFT - with the assumption that good model explanations should be consistent with the explanations from humans who understand the model.
- **Robustness:** Robustness to adversarial perturbation has become an extremely important criterion for applications of machine learning. But, it is nontrivial to find those perturbations. Sharma et al. [31] use the counterfactual method where a counterfactual explanation is a generated point close to an input that changes the prediction and can, therefore, be considered an adversarial example. Using this notion of counterfactuals as adversarial examples, the authors define the Counterfactual Explanation-based Robustness Score (CER-Score) which is the expected distance between the input instance and its corresponding counterfactual. Robust models resist adversarial examples by endeavoring to achieve local robustness at as many points as possible.
- **Complexity:** aims to describe the computational complexity of the method that generates the explanation. The choice to evaluate Complexity using the size of the extracted model is the most accepted. Evans et al. [9] define complexity for only some methods: as the number of splitting points in a decision tree; as the number of rules for Bayesian rule lists; etc.

That said, determining the correct measurement criteria and metric for each case is challenging and remains an open problem in the field. New metrics are proposed and created every month. For example, Lakkaraju et al. [18] argues that to describe the behavior of a given black-box model, it is important to construct an explanation that is not only faithful to the original model but also interpretable. The authors define new metrics as follows:

- **Disagreement:** represents the number of instances for which the label assigned by the black-box model does not match the label assigned by the explanation technique.
- **Unambiguity:** an unambiguous explanation should provide unique rationales for describing how the black-box model behaves in various parts of the feature space. It can be calculated using the overlaps between decision rules in the explanation.

Therefore, Mohseni et al. [23] conducted a review of the evaluation methods used in interpretable machine learning, considering different types of users. . Alvarez-Melis and Jaakkola[1] introduced metrics to quantify the robustness of existing methods and demonstrated that current methods do not perform well according to these metrics. Slack et al. [34] showed that a group of explanation techniques are sensitive to adversarial attacks.

5 DISCUSSION AND CHALLENGES FOR FUTURE RESEARCH

Despite recent progress in interpretable machine learning, there are still some urgent challenges, regarding explanation method design and evaluation. In the previous section, we covered and anticipated some of these challenges, but we certainly did not cover all of the important topics related to interpretable Machine Learning. Based on the performed literature, we capitalized and listed several trends and challenges:

- **Lack of global explanation methods.** While global understanding is particularly important for assessing trust in a model as a whole (before deployment), most of the current research has been devoted to explaining individual predictions. We believe there is an unexplored opportunity in coming

up with explanations that are global in nature and assure trustworthiness before deployment.

- **How to avoid ground truth unjustification?** Post-hoc interpretability approaches are popular for their flexibility as they can be used for any classifier regardless of its training data. However, it also implies that there is no guarantee that the built explanations are faithful to ground-truth data. The risk is providing explanations that are a result of some artifacts learned by the model instead of actual knowledge from the data [19, 20].
- **How can we better evaluate explanations?** Evaluating explanations is, maybe, the most immature aspect of the research on explainablef AI [30]. Setting clear evaluation goals and metrics is critical in order to advance the research on explainability and more efforts are needed in this area. How can we say that an interpretable method is better than another if we do not know why?
- **Can we do better explanations?** Regarding explanations representation, we observed that most studies use numerical and graphical representations, which are complex for non-expert users. Part of the challenge is that the interpretability research field is dominated by machine learning researchers. Mothilal et al. [25] and Lakkaraju et al. [18] argue that users of AI systems should be part of the interpretability designing process from the beginning, and different users need different types of explanations.
  Designing human-centered explanations is a huge field and more dedicated work is necessary to advance in this topic.
- **How does fairness interact with interpretability?** The reviewed literature showed that interpretability methods can be used for bias detection. However, this research area comes with the challenge of identifying how the different fairness measures relate to one another, as well as the extent to which they are compatible or mutually exclusive.
- **How can we build more robust interpretability methods?** A problem of most interpretability methods is their vulnerability to small adversarial perturbations to the input, which incurs a security risk when they are applied to critical areas [34]. Alvarez-Melis and Jaakkola [1] argue that robustness is a key desideratum for interpretability in order to assure trust in Machine Learning models.
- **How to combine and deploy interpretable Machine Learning models?** If we take a close look at the presented approaches, we will find out that while there is some overlap between the various explanation types, for the most part, they appear to be segmented, each one addressing a different question. At this point, we would like to note that there is no established way of combining techniques, so there is room for experimenting and adjusting them, according to the application at hand. This direction could not only help bridge the gap between opaque and transparent models but could also aid the development of state-of-the-art performing explainable models.

6 LIMITATIONS

There are biases related to (i) study selection, i.e., only papers matching the search expression and returned by the searched digital libraries were selected; and (ii) inclusion criteria, i.e., only papers matching the defined criteria were included. Given the novelty of the topic, we believe that articles relevant to the scope of the review may not have been retrieved. We minimized this threat by analyzing the references of the selected studies and recent articles published by their authors. From that, the list of related studies was consulted to verify if they had been analyzed. The vast majority of studies had been read or were already among the group of selected studies. Relevant studies not included were also read in order to improve the review.

7 CONCLUSIONS

The papers considered in this systematic mapping study, as well as its results, are limited by the applied search expression and the research questions. It is not feasible to cover all published papers in this broad field that is exponentially growing. Nevertheless, we believe that our limitations do not have a crucial impact on the results.

This article contributed to one of the Grand Research Challenges in Information Systems: (7) Transparency in Information Systems as Interpretability techniques can serve as a tool to improve transparency and build trustworthy AI systems. Our results showed that the Machine Learning interpretability research field needs to focus more on the comparison of existing explanation methods instead of just creating new ones. Interpretability quality assessment should also include metrics that could help to compare different use cases and methods. Also, we observed that designing human-centered explanation methods has been misrepresented in several of the current approaches. The explanations provided by the different approaches are usually complex for non-expert users. Furthermore, a long-term research direction would be building model-agnostic frameworks capable of recommending the best explanation among the available ones while considering the problem domain, use case, and user.

At the very least, we hoped to introduce some fundamental concepts, and cover several important areas of the field, and show how they relate to each other and challenges still to be addressed. Also, there are many possible directions for future research based on our results. Finally, it is hoped that this study will be a start point for other researchers in the field and help to close some of the discussed interpretability gaps.

REFERENCES

[1] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. *CoRR* abs/1806.08049 (2018), 6. arXiv:1806.08049 http://arxiv.org/abs/1806.08049

[2] 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

[3] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. arXiv:1909.06342 [cs.LG]

[4] Adrien Bibal and Benoît Frenay. 2016. Interpretability of Machine Learning Models and Representations: an Introduction. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Michel Verleysen (Ed.). CIACO, Bruges, Belgium, 77–82. 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2016 ; Conference date: 27-04-2016 Through 29-05-2016.

[5] Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. 2016. Supervised Neighborhoods for Distributed Nonparametric Regression. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 51)*, Arthur Gretton and Christian C. Robert (Eds.). PMLR, Cadiz, Spain, 1450–1459. https://proceedings.mlr.press/v51/bloniarz16.html

[6] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019), 34. https://doi.org/10.3390/electronics8080832

[7] Mark W. Craven and Jude W. Shavlik. 1995. Extracting Tree-structured Representations of Trained Networks. In *Proceedings of the 8th International Conference on Neural Information Processing Systems* (Denver, Colorado) *(NIPS'95)*. MIT Press, Cambridge, MA, USA, 24–30. Available at http://dl.acm.org/citation.cfm?id=2998828.2998832.

[8] Mengnan Du, Ninghao Liu, and Xia Hu. 2019. Techniques for Interpretable Machine Learning. arXiv:1808.00033 [cs.LG]

[9] Benjamin P. Evans, Bing Xue, and Mengjie Zhang. 2019. What's inside the Black-Box? A Genetic Programming Method for Interpreting Complex Machine Learning Models. In *Proceedings of the Genetic and Evolutionary Computation Conference* (Prague, Czech Republic) *(GECCO '19)*. Association for Computing Machinery, New York, NY, USA, 1012–1020. https://doi.org/10.1145/3321707.3321726

[10] Timnit Gebru. 2019. Oxford Handbook on AI Ethics Book Chapter on Race and Gender. arXiv:1908.06165 [cs.CY]

[11] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2019. Explaining Explanations: An Overview of Interpretability of Machine Learning. arXiv:1806.00069 [cs.AI]

[12] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine* 38, 3 (02 Oct 2017), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A Survey Of Methods For Explaining Black Box Models. arXiv:1802.01933 [cs.CY] https://arxiv.org/abs/1802.01933

[14] David Gunning. 2017. *Explainable Artificial Intelligence (XAI)*. DARPA. Available at https://www.darpa.mil/attachments/XAIProgramUpdate.pdf.

[15] Mark Ibrahim, Melissa Louie, Ceena Modarres, and John Paisley. 2019. Global Explanations of Neural Networks: Mapping the Landscape of Predictions. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 279–287. https://doi.org/10.1145/3306618.3314230

[16] Ulf Johansson, Rikard König, and Lars Niklasson. 2010. Genetic Rule Extraction Optimizing Brier Score. In *Proceedings of the 12th Annual Conference on Genetic and Evolutionary Computation* (Portland, Oregon, USA) *(GECCO '10)*. Association for Computing Machinery, New York, NY, USA, 1007–1014. https://doi.org/10.1145/1830483.1830668

[17] Jalil Kazemitabar, Arash Amini, Adam Bloniarz, and Ameet S Talwalkar. 2017. Variable Importance Using Decision Trees. 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., Long Beach, California, USA, 426–435. http://papers.nips.cc/paper/6646-variable-importance-using-decision-trees.pdf

[18] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/3306618.3314229

[19] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations. arXiv:1907.09294 [cs.LG]

[20] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. Unjustified Classification Regions and Counterfactual Explanations In Machine Learning. In *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2019 (Lecture Notes in Computer Science, Vol. 11907)*. Springer, Würzburg, Germany, 37–54. https://doi.org/10.1007/978-3-030-46147-8_3

[21] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why Does My Model Fail? Contrastive Local Explanations for Retail Forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 90–98. https://doi.org/10.1145/3351095.3372824

[22] Andreas Messalas, Yiannis Kanellopoulos, and Christos Makris. 2019. Model-Agnostic Interpretability with Shapley Values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*. IEEE, Patras, Greece, 1–7.

[23] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv:1811.11839 [cs.HC]

[24] Christoph Molnar. 2017. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book

[25] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. https://doi.org/10.1145/3351095.3372850

[26] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, USA.

[27] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. 2018. Model Agnostic Supervised Local Explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 2520–2529.

[28] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-Agnostic Interpretability of Machine Learning. *ArXiv* abs/1606.05386 (2016), 5.

[29] 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

[30] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of User-Centered Explainable AI.

[31] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-Box Models. 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, 166–172. https://doi.org/10.1145/3375627.3375812

[32] Shohei Shirataki and Saneyasu Yamaguchi. 2017. A study on interpretability of decision of machine learning. In *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Boston, MA, USA, 4830–4831. https://doi.org/10.1109/BigData.2017.8258557

[33] Dylan Slack, Friedler, Sorelle A., and Emile Givental. 2020. Fairness Warnings and Fair-MAML: Learning Fairly with Minimal Data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 200–209. https://doi.org/10.1145/3351095.3372839

[34] 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

[35] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard journal of law & technology* 31 (04 2018), 841–887.