Black Box Model Explanations and the Human Interpretability Expectations - An Analysis in the Context of Homicide Prediction

José de Sousa Ribeiro Filho, Nikolas Jorge Santiago Carneiro, and Ronnie Cley de Oliveira Alves

1 Federal University of Pará - UFPA, Belém, Brazil
2 Federal Institute of Pará - IFPA, Ananindeua, Brazil
3 Vale Institute of Technology - ITV DS, Belém, Brazil
*Corresponding Author:
E-mail: jose.ribeiro@ifpa.edu.br
Phone: +55 91 98185-3166

Abstract. Strategies based on Explainable Artificial Intelligence - XAI have promoted better human interpretability of the results of black box machine learning models. This sets a precedent for questioning whether or not human expectations are being met when faced with the explanations of this type of model. The XAI measures being currently used (Ciu, Dalex, Eli5, Lof0, Shap, and Skater) provide various forms of explanations, including global rankings of relevance of attributes, which allow for an overview of how the model is explained as a result of its inputs and outputs. These measures provide for an increase in the explainability of the model and a greater interpretability grounded on the context of the problem. Current research points to the need for further studies (within a specific context/problem) on how these explanations meet the Interpretability Expectations of human experts and how they can be used to make the model even more transparent while taking into account specific complexities of the model and dataset being analyzed, as well as important human factors of sensitive real-world contexts/problems. Intending to shed light on the explanations generated by XAI measures and their interpretabilities, this research addresses a real-world classification problem related to homicide prediction, duly endorsed by the scientific community, replicated its proposed black box model and used 6 different XAI measures to generate explanations and 6 different human experts to generate what this research referred to as Interpretability Expectations - IE. The results were computed by means of comparative analysis and identification of relationships among all the attribute ranks produced, and 49% concordance was found among attributes indicated by means of XAI measures and human experts, 41% exclusively by XAI measures and 10% exclusively by human experts. The results allow for answering questions such as: “Do the different XAI measures generate similar explanations for the proposed problem?”, “Are the interpretability expectations generated among different human experts similar?”,
explanations generated by XAI measures meet the interpretability expectations of human experts?” and “Can Interpretability Explanations and Expectations work together?”, all of which concerning the context of homicide prediction.

**Keywords:** Explainable Artificial Intelligence - XAI · Black Box Model · Human Explainable AI

## 1 Introduction

In recent years, technology has increasingly evolved and allowed intelligent algorithms to be present in our daily lives through solutions to the most diverse types of problems, thus further requiring that machine learning models solve increasingly complex problems requiring characteristics that justify their decision-making capabilities.[51,20].

Computational models based on bagging and boosting algorithms, because they provide high performance and high generalization capacity, are commonly used in computing to solve regression and classification problems based on tabular data. However, these models are not considered transparent algorithms[4] being considered black box algorithms[5] and, therefore, are less used in problems related to sensitive contexts, such as health and safety.[58,34].

By observing the most recent literature on Explainable Artificial Intelligence - XAI[26], the use of black box algorithms in sensitive real-world contexts requires confidence (on the part of the human user) to be gained in the predictions of this type of algorithm. In this sense, different strategies have been developed on two knowledge fronts, namely: one aimed at generating greater explainability of the model itself; the other front with analyses concerning the interpretability of the explanations produced (interpretations made by a human user)[7,41,22].

Black box model explanations are created through Model Agnostic analyses[6] also referred to as Model Inductions[42,25] or even Post-hoc Analysis[7], since in this type of technique only the training data, test data, the model itself and its outputs are used for creating explanations.

The limited understanding of black box models requires the search for methods and tools that can provide information about local explanations — aiming at predicting around an instance through various methods to obtain a local attribute importance ranking[39] — and global explanations — when it is possible to understand the rationale of all instances of the model by generating a global attribute importance ranking[39,24] — as a means of making interpretable, and thus more reliable, decisions[25].

---

4 Transparent Algorithms: Algorithms that generate explanations for how a given output was produced. Examples: Decision Tree, Logistic Regression and K-nearest Neighbors[7].
5 Black Box Algorithms: Machine learning algorithms that have classification or regression decisions that are hidden from the user[10].
6 Model-Agnostic: meaning that it does not depend on the type of model to be explained[42].
Efforts have been made in the Explainable Artificial Intelligence - XAI area regarding the development of different measures to explain black box models, even after their training and testing process [7]. Thus, measures such as Ciu [19], Dalex [11], Eli5 [35], Lofo [13], Shap [37], and Skater [3] have emerged to provide the creation of model-agnostic explanations. Each of these tools is capable of generating explanations using different techniques, but a fact they have in common is that they all generate global attribute importance rankings related to the explanation of a model.

The explicabilities generated by each of the above-mentioned XAI measures, as seen in [48], are ranks that can have greater similarity to one another, even if they stem from different XAI measures, in cases where the complexity of the model (dataset and algorithm) is lower; however, for models with higher complexities, these ranks can exhibit greater differences from one another, as seen in [48], factors that undermine the confidence gain of the model’s end users.

Seeking to minimize the problems presented so far, besides the development of XAI measures and studies on human interpretability, another initiative that has stood out and provided an interesting confidence gain in black box models is the inclusion of human interactions during the process of creating the explanations of the models [50,43,54,59], as this secures the control of explanations as well as a greater scope thereof linked to interpretations of the context based on human and machine analysis.

By that, given the context of homicide prediction and the various research fronts involving explainability, interpretability and human interactions in the black box opening process, the following questions arise: “Do the different XAI measures generate similar explanations for the proposed problem?”, “Are the interpretability expectations generated among different human experts similar?”,”Do the explanations generated by XAI measures meet the interpretability expectations of human experts?” and “Can Explanations and Expectations of Interpretability work together?”.

By seeking to answer these questions, an experiment was developed that uses the machine learning model of homicide prediction advocated in [47], and from this, the 6 explicability rankings were generated by means of XAI metrics, as in [48], and 6 ranks of interpretability expectations generated by different human experts. Then, comparisons and identification of existing relationships between all pairs and sets of ranks created were performed to find the desired answers. Finally, the generated ranks were combined into a single overall rank by means of a technique proposed hereby based on the results of the explanations of the XAI measures and interpretability expectations that were found.

The main contributions of this research are as follows:

- To shed light on the discussion regarding the similarity of explanations generated by XAI measures and their interpretability, focusing on the specifics of the context-sensitive problem - predicting homicides - in order to measure
whether the XAI measures explain the model as expected by human experts, even if to this end some of the results cannot be easily replicated for other regions and other contexts, given the level of the analyzed specificities of the context concerned.

- To propose the concept of Interpretability Expectation, which in general terms is the interpretability expected by an expert of a real-world problem based on their knowledge of the problem and the working principle of the machine learning model being analyzed, thus allowing this concept to be used and replicated by the XAI community in other contexts/problems.

- To describe how the results of the Interpretability Expectation can be combined with results of XAI measures, which are based on global attribute importance rankings, in order to build a Collaborative Explanation of the model using human expert knowledge and different XAI measures, i.e. human and machine. This technique may well be used in other contexts/problems.

- To show the overall methodology developed by this study as a deliverable, as it promotes data used, code developed, results collected, and the repositories created, in accordance with the Fair Guiding Principles for scientific data management and stewardship.

2 Background

2.1 Explainability and Interpretability

The concepts of explainability and interpretability in machine learning are considerably close and even complement one another [7,42]. Therefore, it is of utmost importance that they are presented and differentiated.

Explainability is associated with the explanatory interface between a computational model and a human, which aids in the decision-making process as it seeks to make the model understandable [7,42].

Interpretability is the ability to provide meaning in terms that are understandable to a human being, or even the attempt to interpret an explanation [41,42,22].

Based on these two concepts, which are widespread in the area of machine learning, it is understood that in a practical way explainability seeks to create subsidies that explain the black box model in a technical manner, whereas interpretability seeks to give meaning to the explanations created for a human user, such meaning being based on the context of the problem and the knowledge of the individual [12,47].

Both explainability and interpretability of models are fundamental pieces in the decision-making process, as they provide the end user with support in detecting various problems or even biases in the data being used by the model [41,53].

As mentioned above, it is not possible to conduct a study involving analysis of explainability and interpretability of computer models without considering the

---

8 Explain the black box model: Also known as the process of “opening the black box”.
specific context/problem in which they are embedded and the human factors as well [7].

In this sense, this research focuses on a single specific problem to perform its analysis. In addition to this and also to issues of time and cost feasibility, the context of homicide prediction was chosen.

Therefore, it can be assumed that explainability and interpretability allow the generation of confidence, understanding, and fairness to black box machine learning models. In the studies and experiments described herein, the main focus is on the explainability of each generated model and its relationship to the interpretabilities (in this case, expectations) generated by humans in the context of crime prediction.

2.2 Measures of Explainable Artificial Intelligence

Despite the wide applicability of ensemble-based algorithms (such as: bagging and boosting) in real-world problems, this group of algorithms has a smaller quantity of explainability measures and benchmark studies when compared to the current moment of studies concerning the explainability of neural networks [36][1]. That is, there are still few model-agnostic XAI techniques.

Currently, there are different types of XAI techniques aimed at generating global model-agnostic explanations, namely: Partial Dependence Plots - PDP, Accumulated Local Effects - ALE, Feature Interaction, Functional Decomposition, Permutation Features Importance, Global Surrogate, and Prototypes and Criticisms [42].

Based on these types of measures listed above, this research focuses on two types only: Permutation Feature Importance and Global Surrogate, since these two categories are capable of generating ranks and feature-compatible implementations. Based on the above, this research conducted a bibliographic and practical survey (development) on the main existing XAI measures, specifically aimed at generating model-agnostic global explainability ranks that support tabular data. As a result, a total of six tools were selected, with libraries updated and compatible with each other, with current Python development platforms and Scikit-Learning algorithms [2]. These tools are: CIU [19], Dalex [8], Eli5 [35], Lofo [13], SHAP [37] e Skater [8].

The measures referenced herein generate explanatory ranks based on the same previously-trained machine learning models, manipulate their inputs or/and produce new intermediate models (copies). Therefore, a comparison of the generated explanatory ranks is fair and feasible.

It should be noted that during the initial analyzes and executions, as occurred in [48], this research found XAI measures with incompatibilities at the level of libraries and dependencies, making it impossible to use and compare some measures. For this reason, some XAI means were not used for this research (for example: Alibe-ALE [4], Lime [49], Ethical ExplainableAI [18], IBM Explainable AI 360 [5], and Interpreter ML [44]).

The Contextual Importance and Utility - CIU is a XAI measure based on Decision Theory [33] that focuses on serving as a unified metric of model-agnostic
explainability based on the implementation of two different indexes: Contextual Importance - CI and Contextual Utility - CU. As verified in preliminary tests carried out by this research, these two indices generate equal ranks. Thus, it was decided to use the CI[19].

The measure Dalex is a set of XAI tools based on the LOCO (leave one covariate out) approach and can generate explainabilities from this approach. This measure receives the model and the data to be explained, calculates model performance, performs new training processes with new generated data sets, and makes the inversion of each attribute of the data in a unitary and iterative way, measures what attributes are important to the model, evaluates its performance obtained according to the inversions of the attributes[6,11].

A little less popular but very powerful is the Leave One Feature Out - Lofo, a XAI measure with a similar proposal to that of Dalex, but no attribute inversion is performed here, because in the Lofo metric the iterative step is based on iterative removal of the attributes to find its global importance to the model. This measure also analyzes the performance of the model[13].

A very popular and quick measure to be performed, the Explain Like I’m Five - Eli5 is a tool that helps explore machine learning classifiers and explains the predictions thereof by assigning weights to decisions, as well as exporting decision trees and presenting the importance of the attributes of the model submitted to the tool. In its documentation, it is possible to see that this measure uses only training and test data to create the explanations, which is curious, given that the weights it extracts are based on a tree algorithm[35].

One of the most popular and currently used measures is the SHapley Additive exPlanations - SHAP, proposed as a unified measure of attribute importance that explains the prediction of an instance \( X \) from the contribution of an attribute. What makes this measure different from the others is the calculation explanation index based in the game theory of Shapley Value[50,40,37].

Last but not least is the measure Skater, a set of tools capable of generating ranks of the importance of model attributes, differing from the other measures to calculate its explanation index based on Information Theory[3], through measurements of entropy in changing predictions through a disturbance of a certain attribute[9].

A general comparison of the main characteristics of the measures that meet the requirements of this research is shown in table 1.

| XAI measure | Author | Base algorithm | Explanation Technique | Global explanation (by rank) | Local explanation | API compatible |
|-------------|--------|----------------|------------------------|-----------------------------|-------------------|----------------|
| CHU | [19] | Decision Theory | Multiple Criteria Decision Making | Yes | No | Yes |
| Dalex | [8] | Leave-one covariate out | Permutation of Feature | Yes | Yes | Yes |
| Eli5 | [35] | Assigning weights to decisions | Permutation of Feature and Mean Decrease Accuracy | Yes | Yes | Yes |
| Lofo | [13] | Leave One Feature Out | Permutation of Feature | Yes | No | Yes |
| SHAP for Tree | [40] | Game Theory | Permutation of Feature | Yes | Yes | Yes |
| Skater | [3] | Information Theory | Permutation of Feature | Yes | Yes | Yes |
All the tools described above are capable of generating different types of explanations (instance-level explanation, dependent plot, impacts on the model output, similarity of instances, among others), which go beyond global explanations based on attribute ranks; therefore, the focus of this article is to compare this most basic unit of model explainability, i.e., the global importance rank.

Even knowing that each of the XAI measures used in this research are based on different algorithms and methodologies, a comparison of their results is fair. Since it will only compare the basic structure of your explanations based on attribute rank.

3 Materials and Methods

This research performed analyses involving the machine learning model (Figure 1 (1)) as advocated in [47], duly trained (Figure 1 (2), (3) and (4)), together with executions of the XAI measurements (Figure 1 (5)) and queries with human experts (Figure 1 (6)), so as to generate the explicability ranks (Figure 1 (7)) and interpretability expectations (Figure 1 (8)), thus allowing for performing analyses from three different perspectives (Figure 1 (9), (10) and (11)). All the steps in the processes involved in the developed methodology are presented in the following topics.

Fig. 1. Visual scheme of all steps and processes performed by the proposed methodology.

All information regarding the reproducibility of the methodology described here can be accessed through the link in the “Data Available” area.
3.1 Prediction of homicides model

The machine learning model used in the analyses performed herein has already been duly validated in [47], being developed on the basis of the data provided by the Office of Intelligence and Criminal Analysis - SIAC of Pará State, Brazil. Such data refer to the police reports registered during the years 2016 to 2018 in the city of Belém, Pará State.

The City of Belém, capital of the state of Pará, Brazil, the scenario of this study, is a city with 1,393,399 inhabitants, according to the last census performed in 2010, and has a human development index of 0.746 [29]. This city has institutions that carry out several programs in the area of public security, and in recent years there has been a decrease in cases of violent crimes.

Nevertheless, Belém featured the third worst homicide rate among all Brazilian capitals (74.3) in a 2017 study [30]. Thus, the data from the police records in this city prove to be interesting sources of information for the herein study.

In a recent survey, Pará State ranks together with eight other Brazilian states as having incomplete data regarding homicide cases and solution thereof [45]. Thus, the importance of the study conducted in [47] is hereby emphasized, since the developed model proved capable of carrying out the crime prediction process, even under different circumstances. In the future, the defended methodology may be able to help solve problems related to missing data on violent crimes.

As advocated in [47], the machine learning model proposed hereby is based on the dynamics of how different crimes that occur in a city can be explained by different theories in the area of Social Sciences, such as those related to a single person, called “Theory of Understanding the Motivations of Individual Behavior”, and also from the “Theory of Associated Epidemiology”, related to the study of how criminal behavior is distributed and displaced in time and space of a given location [16,12].

This second theory, presented above, supports the way in which this research understands the relationships between different types of crimes, that is, through a deep analysis of the relationships of the most different crimes and their time and space, the possibility of predicting a specific crime is suggested [47].

Thus, for the construction of the model, several transformations were performed on the original database of police reports, in order to allow the quantity of each crime occurring in years, months and neighborhoods to be correlated to the binary class referring to the occurrence or not of homicides in the following month, thus making it possible to analyze how different crimes are dynamically related to homicides in the city of study, since the occurrence of specific crimes is related to a context of the conflict between individuals, and one offense may influence homicides [21][16][14].

As seen in [47], the data used as inputs for machine learning models are completely generic, since they are made up of the month variable (1 variable) along with the different numbers of crimes that occurred in neighborhoods in a specific month (34 variables), which makes this methodology of using criminology data for the prediction of homicides easily replicable to other cities that have the
same data. Only at the level of curiosity, Figure 2 shows the information gain calculated for the data of the proposed system.

![Informatio Gain by feature](image)

**Fig. 2.** Rank of the 10 attributes with the highest values of information gain in the dataset of the crime prediction model.

It is also noteworthy that all records of different types of crimes (34 types in total) were selected by a public security specialist. By that, only crime types that can be directly or indirectly linked to homicides are found in the final dataset [47]. More details about the dataset and analysis of this study can be found at: [https://github.com/josesousaribeiro/Pred2Town](https://github.com/josesousaribeiro/Pred2Town).

According to [47], eleven (11) different algorithms were tested for the homicide prediction problem. However, the results with the best performance and stability were only found by the Random Forest - RF algorithm [9], thus representing the best machine learning model developed. For this reason, the methodology applied in the study presented herein will use this model for the analyses presented in the topic Materials and Methods.

### 3.2 Rank by XAI measures

As seen earlier, it is known to this research that the XAI measures used (Chu, Dalex, Eli5, Lof0, Shap and Skater) have different base algorithms that act in creating the explanations of black box models. However, it is emphasized that all
these measures are compatible with each other and are comparable, since they all generate the basic rank structure as an explanation for an analyzed model.

The XAI averages can create ranks of explanations based on the importance that a particular attribute has for the model as a whole. In general, these ranks are ordered based on a score, defined by the measure and calculated for each attribute, thus enabling the generation of ranks for a maximum limit of attributes $n$ unknown to this research [42].

A sensitive, but of utmost importance, detail should be observed and properly described regarding the generation of the ranks, since an XAI measure can generate a rank of explainability with more than one attribute with the same score. By that, when generating each of the XAI measure ranks, each attribute with the same score was ordered by its label (attribute name), thus ensuring a consistent comparison between the measures.

As above, the 6 XAI measures were executed for the homicide prediction model, and thus 6 explicability ranks were created containing the order of importance of the 35 attributes in the model.

### 3.3 Rank by Human Specialists

This research contacted the people in charge of the Office of Intelligence and Criminal Analysis - SIAC of Pará State to query a total of 6 experts in the field of criminology and thus be able to create the explainability ranks. The answer was the assignment of 6 specialists in the field, with different profiles: Public Security Agents, Military Police Officers, Criminal Data Collection Technicians and Data Collection/Analysis Managers. It should be noted that none of the specialists has a computer science background, specifically.

Each of the specialists mentioned above was assigned the task of creating a rank of importance of attributes, but in order for this to become humanly feasible, instead of ranking the 35 input variables of the model, they were asked to rank 10 attributes only, since a total of 10 attributes is more easily sortable or even classifiable and conforms to the limits of an ordinal measurement scale for this task [57].

During the rank, in parallel, each expert assigned an ordinal value from 1 to 10 to each of the 10 attributes (in this case, the closer to 1 the attribute is the more important it is). This enabled the creation of explainability ranks with 10 most important sub-deemed attributes for the homicide prediction process.

The form applied to each human expert can be viewed at: — https://sites.google.com/view/pred2town/in%C3%ADcio.

As described above, having in hand the ranks of XAI measures as well as the ranks generated by human experts, this research focused on working on Comparisons, Relationships and Rank Combinations analysis to understand how explanations and interpretabilities are generated, as well as to propose a way to combine the generated ranks into a single consolidated result.
3.4 Rank Comparision and Relationship

Initially, this research elected to use rank coefficients, as is the case of Spearman’s Coefficient and also that of Kendall [17], as a way to measure similarities and correlations between the different pairs of generated ranks. However, it was verified that, by taking into account only the first 10 attributes of each rank, data loss in some comparisons is generated and this impairs the use of the rank coefficients mentioned above [38].

Thus, a decision was made to use a simpler similarity check for each pair of ranks by counting the attributes existing simultaneously in the two ranks (without considering their position), that is, each rank is now considered as a set and to check their similarity the intersection between all pairs of sets generated through the basic set theory is calculated [31].

This research understands that analyzing each attribute rank — generated as explainability ranks for each XAI measure and human experts — by means of set theory allows a transparent analysis as to the intersection between the sets and the subtraction between them. Therefore, this metric was used to compare the ranks and to understand their relationships.

3.5 Interpretability Expectation

The term interpretability expectation is proposed hereby in a generic way to refer to what is expected by the human expert as interpretations of the prediction model, this term is inspired by research in the area of XAI that questions the importance of human interaction with the explanations generated by XAI measures applied to black box models [56,55,13] and how the interpretability of explanations of machine learning models are related to the expert’s expectation of a problem by confronting their previous experiences and also the confidence in the model being analyzed [32,27,25,23].

Thus, to generate the expectation of interpretability, the expert analyzes the machine learning model input data (attributes) and its performance (for example, accuracy), considers their pre-knowledge of the proposed problem (context) by re-signifying the use of each attribute, and finally performs the creation of what is expected as a possible explanation to the model (for example, rank of importance of attributes) based on their interpretation of what is considered more consistent for the problem.

In other words, the interpretability expectation of the machine learning model is an explanation created by a human expert who is based upon their interpretation of the context.

It is important to point out that, in order to carry out the process of creating the interpretability expectation advocated above, this research deems as important that the expert has no previous contact with the results of the explanations generated by the XAI measures, so as to avoid the creation of biases.

Another important highlight is that the interpretability expectation does not emerge as a universal explanation proposal for a model, but can be easily
complemented by explanations generated through XAI measures, given the capability of these tools. In this sense, this research seeks to introduce a proposal to unify global attribute importance rankings generated from XAI measures to the rankings obtained from the application of the interpretability expectation.

This technique aims at providing the inclusion of the human being in the loop of the process of opening the black box models, since in a next step it can enable the combination of results obtained by XAI measures and human experts.

3.6 Combining Explicability and Interpretability Expectation

As seen in the topic above, once in possession of XAI measures ranks and interpretability expectations ranks, comparative analyses can be performed by using set theory fundamentals to obtain results regarding comparisons of ranks and relationships between sets of analyzed ranks. However, these two techniques do only focus on the differences and similarities existing within all the ranks and sets of ranks under analysis.

Another interesting and innovative way to use the results generated both by XAI measures and by human experts is to combine (or even include) the explainability results with the interpretability results, with the purpose of guaranteeing the human element in the process of opening the black box models, called by this research Combining Explanations and Interpretations - ConeXi.

Based on the simple concepts of sigma notation, commonly used in mathematics [10] and algorithm complexity [52], the notations 1 and 2 are achieved, which together represent the mathematical process performed to calculate scores for each feature and thus combine the ranks of explainability and interpretability, ultimately aiming at creating a single overall ordered rank where features closer to 1 have higher explainability and interpretability.

Notations 1 and 2 represent a $S$ point counting system applied to each feature in the ranks. For this, you need to consider the notations and their variables:

\[
S = \sum_{i=m}^{n} f_i
\]  

where $n$ is the number of features in each rank (in this study, it is equal to 10). $m$ is the initial index of the iteration (equal to 1). Finally, $f_i$ is the number of times that a certain attribute appeared in a rank at a position equal to or lower than $i$.

For better understanding, $f_i$ can be represented as a second sum represented by 2

\[
f_i = \sum_{j=q}^{r} p_j \ast w_j
\]  

where $r$ represents the number of ranks existing in the analysis (in this study, it is equal to 12). $q$ is the initial index of the iteration (equal to 1). Finally, $p_j$ is
equal to 1 if the feature concerned is in a position equal to or lower than \( i \) in the rank \( j \), otherwise it is equal to 0. Lastly, the value of \( w \) is the weight that will multiply the score of a specific rank, thus maximizing or minimizing it (herein using \( w_j = 1 \)), a value that can be manipulated at the end of the process by the human user who seeks to explain the model.

After calculating the \( S \) value for all the features in the analyzed ranks, one can simply sort the \( S \) values in descending order and thus find an overall rank that is the combination of the explainability ranks and the interpretability ranks, as shown in Figure 6 further ahead.

4 Results and Discussion

The results of this paper are divided into three different analyses and seek to answer the questions posed at the beginning of this article, Figure 3.

| Analysis | Questions |
| --- | --- |
| Comparison between Explanations and Expectations of Peer Interpretability | "Do the different XAI measures generate similar explanations for the proposed problem?" |
| Relationship between Explanations and Interpretability Expectations | "Are the interpretability expectations generated among different human experts similar?" |
| Combination of Explanations and Interpretability Expectations | "Do the explanations generated by XAI measures meet the Interpretability expectations of human experts?" |
|  | "Can Explanations and Expectations of Interpretability work together?" |

Fig. 3. Survey analyses vs. questions to be answered.

4.1 Comparison between Explanations and Expectations of Peer Interpretabilities

Peer Comparison of Explanations and Interpretability Expectations is defined as all the comparative analyses performed between the explanations generated individually by each of the XAI measures and individual interpretability expectations of each human expert.

To this end, a general graph was built showing the arrangement of the different amounts of attributes existing at the intersections of all the pairs of sets generated, Figure 4. Considering that such comparisons involve: Explanation Measures only (EM vs EM), Interpretability Expectations only (IE vs IE), and both Explanation Measures and Interpretability Expectations (ME vs IE).

Figure 4 summarizes all the comparisons that were performed, since the axis \( y \) shows the different amounts of attributes existing at the intersection between each of the pairs of sets analyzed and arranged on the axis \( x \).
From the results of the ME vs ME comparisons (explanation measures only), it is noticeable that the medians of the intersections showed a slight drop (mean equal to 6.07). However, this is directly influenced by a higher variance that is noticeable in the graph, since the comparison “ME: Eli5 vs ME: Skater” had the highest number of attributes in its intersection (total of 9 attributes), but “ME: Dalex vs ME: Lofo” and “ME: Ci vs ME: Lofo” have the lowest number of attributes in their intersections (total of 3 attributes in each).

The results above show indications that the dataset being analyzed has considerable complexity, given the existence of different explainability ranks as results of the execution of the different XAI measures. Similarly to what was seen in [48].

Thus, for the context proposed herein, the different XAI measures agree with one another in most of the indications of attributes that explain the computational model, answering the first question posed by this research “Do the different XAI measures generate similar explanations to one another for the proposed problem?”.

The results given by the IE vs IE comparisons (interpretability expectation), Figure 4 - red circles, show that the 6 different human experts agree with each other with an average of 6.73 attributes. Upon closer inspection of the chart, it can be seen that the pairs of interpretability expectations that are generated agree on at least 5 attributes, and a maximum of 8 attributes. This fact shows that the specialists have their expectations minimally aligned with one another, thus answering the second question posed by this research “Are the interpretability expectations generated among different human experts similar?”.

Fig. 4. Summary of the quantities of intersections between pairs of different sets of explanations. Axis y shows the quantity of each intersection and axis x shows each pair of compared set. The colors red, green and blue refer to three different types of comparisons, respectively: IE vs IE, ME vs ME and ME vs IE.
As additional results, specific to some of the XAI measures that were analyzed, it is noteworthy that the CI (of the CIU measure) and the Shap, were those having the highest number of attributes in common with the interpretability expectations of the experts, a total of 6 attributes, (Figure 4 “ME: ci vs IE: human 5”, “ME: shap vs IE: human 6”, and “ME: ci vs IE: human 6”). Therefore, these are the measures that showed the closest results to that of at least one of the experts.

Conversely to the previous analysis, it should be noted that the Lofo measure showed results that were the furthest from the experts’ interpretability expectations, presenting 2 attributes in common with the human experts (Figure 4 “ME: lofo vs IE human 1”, “ME: lofo vs IE human 2”, “ME: lofo vs IE human 3”, e “ME: lofo vs IE human 5”).

In advance of some results that will be better addressed in the following topic herein, still with regard to the Figure 4, the results of the ME vs. IE comparisons (measures of explainability vs. interpretability expectations only) show the opposite of what was observed in the previous two analyses, as now the mean of the intersections show an average of 4.03. That is, in most comparisons the intersections between explanations of XAI measures and expectations of interpretabilities showed more disagreement than agreement with each other, thus providing a preview of what could be the answer to the third question posed hereby “Do the explanations generated by XAI measures meet the interpretability expectations of human experts?”. However, this result is overshadowed due to the perspective presented below.

This last result described above, in the context of the proposed crime prediction, portrays a sensitive problem usually faced by managers who use Artificial Intelligence to support decision making, since, as can be seen, there are expectations of specific interpretability generated by human experts in the public security area that are different from the explanations generated by specific XAI measures.

In the following topic, a new approach for identifying the relationship between EM and IE is addressed herein. By that, the answer to the last question posed in this research is reached.

4.2 Relationship between Explanations and Interpretability Expectations

As a relationship between Explanations and Interpretability Expectations, comparative analyses are defined as performed by creating two main sets - one set containing all the attributes indicated by the measures of explanations (called SME) and another set containing the attributes referring to interpretability expectations (called SIE) - substantiating the results on the context of the proposed problem.

As previously mentioned, it is important to emphasize that at no time does this research intend to question the attributes pointed out by specialists in the area of security, because each of these attributes was consciously indicated by
professionals who deal directly or indirectly with data related to the confrontation of violent crimes in the city of Belém, in the state of Pará.

In this regard, two large sets, SME and SIE, were created without the presence of attribute repetitions, and the intersection between these two large sets were calculated, the result being shown in Figure 5.

Figure 5 clearly shows that the SIE set has a total of 16 attributes and the SME set has 24 attributes. These two sets have an intersection of 13 attributes, so they agree with an approximate percentage of 49% of the total of attributes.

The results above show that despite the high number of attributes indicated by the explanability measures (24 attributes in total), there are attributes from the set of interpretability expectations that were not related. So, this shows that approximately 82% of the interpretability expectations of the experts involved in this research were met (considering 16 SIE attributes = 100%), thus answering the third question posed earlier in this paper: “Do the explanations generated by XAI measures meet the interpretability expectations of human experts?”.

The high number of attributes that exist exclusively in SME (11 in total) demonstrates that the XAI measures point to attributes that are not being considered as priorities to human experts, thus, indicating that experts can explore even more explanatory variables of the model.

Table 2 shows aspects related to the context of homicide predictions to the results of this research.

This research understands that the results presented in the set of interpretability expectation, \((SIE - SME) + (SIE \cap SME)\) refer to Table 2 are attributes strongly related to the proposed problem and to human interpretation, since they were directly indicated by experts who deal with crime-related data in their daily lives and have relevant knowledge in the public security area due to their professional backgrounds. Therefore, it is not for this research to question
Table 2. Attributes in the SIE and SME

| Set               | SIE - SME | SIE ∩ SME | SME - SIE |
|-------------------|-----------|-----------|-----------|
| Quantity of attributes | 3         | 13        | 11        |
| Name of the attributes |           |           |           |
| - Threat;          |           |           |           |
| - Physical Assault;|           |           |           |
| - Family conflicts;|           |           |           |
| - Neighboring conflicts;|   |           |           |
| - Missing person;  |           |           |           |
| - Rape;           |           |           |           |
| - Homicide;       |           |           |           |
| - Personal Injury;|           |           |           |
| - Month;          |           |           |           |
| - Robbery;        |           |           |           |
| - Theft of vehicles;|     |           |           |
| - Drug trafficking;|           |           |           |
| - Come to blows.  |           |           |           |
| - Embezzlement;   |           |           |           |
| - Desertion.      |           |           |           |
| - Libel;          |           |           |           |
| - Damage;         |           |           |           |
| - Damage in traffic;|   |           |           |
| - Defamation;     |           |           |           |
| - Larceny;        |           |           |           |
| - Theft;          |           |           |           |
| - Slander;        |           |           |           |
| - Traffick injury;|           |           |           |
| - Other atypical events. |       |           |           |

the applicability in the context or even the importance of such attributes, considering these attributes as the ideal of model explanations through the human perspective.

The results presented in the column SME - SIE point to evidence of the existence of a high number of crimes (11 crimes in total) that, according to the proposed model and the explanations that were generated, can explain the process of homicide prediction and can be considered and analyzed by specialists in the field of criminology, so as to allow for a better and broader understanding of the dynamics of crime occurrences in the city under study and new interpretations of the model.

4.3 Combination of Explanations and Interpretability Expectations

As a Combination of Explanations and Interpretability Expectations, a set of analyses are defined that seek to facilitate the combination of the different ranks existing in this study. To this end, this research applied the proposed ranking combination presented earlier to the collected results, and thus constructed an overall ranking visualization that demonstrates the combination of efforts of XAI measures with human experts, Figure 6.

The aforementioned illustration shows the dynamics of the positions of each attribute indicated by the XAI measures (heatmap green) with the attributes indicated the expectations of interpretability (heatmap red) of the experts. Still in Figure 6 there is part referring to iterations (heatmap blue) calculate the features points in the proposed algorithm, and finally the S score is calculated in the last heatmap (heatmap yellow).

Building up Figure 6 required a sorting of the lines according to the values of S, as seen in sigma notations 1 and 2 shown earlier. Therefore, at each iteration of i, the number of times a given attribute appears in the ranks in a position i ≤ x was added up, where x varies from 1 to 10, Figure 6 Iteration table to the right.
In the next step, the Sum of Points, the scores of each attribute are added in $S$. Finally, they are sorted in descending order (Figure 6 Iteration table, column $S$).

By this, Figure 6 represents an overall ranking that explains the model, and is the result of combining the results of the XAI measures together with efforts of human experts. However, it should be noted that through simple changes in the methodology of rank matching, sigma notations 1 and 2 shown previously can be multiplied to the values of $p_j$ by different weights $w_j$, and thus maximize or minimize specific results for given ranks, therefore, for example, giving greater weight to human expert rankings or even to specific XAI measure rankings.

The results presented in the herein study are indicative that both current XAI measures can be improved to come even closer to how human experts explain and interpret the model and the context of the problem. But, also, the results show that experts can consider the explanations generated by XAI measures and enrich their understandings about the context/problem and the analyzed model concerned. Both results can be used in a single Collaborative Explanation, thus answering the third and last question posed hereby: “Can Explanations and Expectations of Interpretability work together?”

Therefore, according to all the results presented herein, it is considered that a Collaborative Explanation between Man and Machine may be the best way to succeed in the opening process of black box models, as even with the great potential and analytical capacity of XAI measures, the interpretability expectations represent results that complement the results pointed out by XAI measures, thus providing new inputs for discussions concerning the context to which the machine learning model is inserted.

### Fig. 6. Heatmap indicating the position (numbers 1 to 10) of each model input attribute in a specific rank (columns).

| Attribute | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|-----------|---|---|---|---|---|---|---|---|---|----|
| Homicide  | 10| 7 | 6 | 3 | 4 | 5 | 2 | 1 | 9 | 8   |
| Assault   | 9 | 7 | 3 | 5 | 2 | 4 | 1 | 10| 8 | 10  |
| Theft     | 8 | 4 | 5 | 1 | 2 | 3 | 10| 9 | 10| 10  |
| Robbery   | 10| 7 | 8 | 5 | 4 | 3 | 2 | 1 | 10| 10  |

In the next step, the Sum of Points, the scores of each attribute are added in $S$. Finally, they are sorted in descending order (Figure 6 Iteration table, column $S$).

By this, Figure 6 represents an overall ranking that explains the model, and is the result of combining the results of the XAI measures together with efforts of human experts. However, it should be noted that through simple changes in the methodology of rank matching, sigma notations 1 and 2 shown previously can be multiplied to the values of $p_j$ by different weights $w_j$, and thus maximize or minimize specific results for given ranks, therefore, for example, giving greater weight to human expert rankings or even to specific XAI measure rankings.

The results presented in the herein study are indicative that both current XAI measures can be improved to come even closer to how human experts explain and interpret the model and the context of the problem. But, also, the results show that experts can consider the explanations generated by XAI measures and enrich their understandings about the context/problem and the analyzed model concerned. Both results can be used in a single Collaborative Explanation, thus answering the third and last question posed hereby: “Can Explanations and Expectations of Interpretability work together?”

Therefore, according to all the results presented herein, it is considered that a Collaborative Explanation between Man and Machine may be the best way to succeed in the opening process of black box models, as even with the great potential and analytical capacity of XAI measures, the interpretability expectations represent results that complement the results pointed out by XAI measures, thus providing new inputs for discussions concerning the context to which the machine learning model is inserted.
5 Final Considerations

However, in view of the sensitive context of the homicide prediction problem in the real world, by analyzing the generated rankings through comparisons and relationships, the expectations of interpretability were noted not to have been fully met at any time, thus revealing the need to have XAI measures that better fulfill the expectations of human interpretability while pointing out attributes that explain the model through new perspectives for criminal experts in order to have a better balance on what is already expected as interpretability and what is still unknown to the human user.

The proposed methodology of combining ranks of interpretability expectations and explainability seems promising, since it allows machine results and human results to be combined and thus tends to overcome the differences identified between these two groups, but this technique is limited to interpretability and explainability based on structured ranks.

These conclusions portray the difficulty in explaining the decisions of black box algorithms in actual scenarios and highlight the need for further studies and metrics in this domain specifically, and highlight the need to have the human being in the loop of the explanation process of machine learning models in order to build a collaborative explanation between man and machine.

6 Future works

In view of the foregoing, in the next steps of this research, validating the concept of interpretability expectation for other contexts is intended in order to expand the understanding thereof to the XAI community. Also, proposing an explainability measure that provides a new way to explain machine learning models is intended as based mainly on data complexity through new techniques grounded on Item Response Theory - IRT to add efforts to the existing XAI measures project.

7 Acknowledgements

The researchers hereof would like to thank the Office of Intelligence and Criminal Analysis - SIAC, in the city of Belém, Pará State, which provided its specialists to voluntarily participate in this study.

Funding

This research was not funded.

Conflicts of interest/Competing interests

It is declared that there are no conflicts of interest between the authors and their institutions belonging to any part of this research.
Ethics approval

There is no need for the approval of a research ethics council, because the personal identity of the consulted experts is not presented, because they work only with the quantitative aspects of their answers and because the data consulted are in the public domain obtained through request to the responsible institution.

Consent to participate

All individuals involved (researchers and volunteers) in this research declare that they are aware of its nature and seriousness, together with its importance to society.

Consent for publication

All individuals and institutions involved in the research in question are in agreement with the publication of this article in the journal.

Availability of data and material

All data used in this study are available at:

- https://github.com/josesousaribeiro/Pred2Town-and-XAI

Code availability

All code used in this study are available at:

- https://github.com/josesousaribeiro/Pred2Town-and-XAI/blob/main/XAI_Pred2Town.ipynb

Authors’ contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by José Ribeiro, Nikolas Carneiro and Ronnie Alves. The first draft of the manuscript was written by José Ribeiro and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.
1. Ethical ML. https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models-and-datasets, Last accessed 20 Apr 2021.

2. Scikit-learn. https://scikit-learn.org/0.22/, Last accessed 21 Nov 2020.

3. Skater. https://oracle.github.io/Skater/overview.html#skater, Last accessed 21 Jan 2021.

4. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(4), 1059–1086 (2020)

5. Arya, V., Bellamy, R.K., Chen, P.Y., Dharanidar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilovic, A., et al.: AI explainability 360: An extensible toolkit for understanding data and machine learning models. J. Mach. Learn. Res. 21(130), 1–6 (2020)

6. Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. arXiv:2012.14406 (2020), https://arxiv.org/abs/2012.14406

7. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Bamba, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115 (Jun 2020), https://doi.org/10.1016/j.inffus.2019.12.012

8. Biecek, P.: Dalex: Explainers for complex predictive models in r. Journal of Machine Learning Research 19(84), 1–5 (2018), https://jmlr.org/papers/v19/18-416.html

9. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

10. Bronshtein, I.N., Semendyayev, K.A.: Handbook of mathematics. Springer Science & Business Media (2013)

11. Burzykowski, P.B.a.T.: Explanatory Model Analysis. https://ema.drwhy.ai/

12. Cano, I., Soares, G.D.: As teorias sobre as causas da criminalidade (2002)

13. Çayır, U., Yenidoğan, I., Dağ, H.: Use case study: Data science application for microsoft malware prediction competition on kaggle. Proceedings Book p. 98 (2019)

14. Cerqueira, D., Lobão, W.: Determinantes da criminalidade: arcabouços teóricos e resultados empíricos. Dados 47, 233–269 (2004)

15. Cortez, P., Embrechts, M.J.: Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). pp. 341–348. IEEE (2011)

16. Cressey, D.P.: Crime: Causes of Crime in International Encyclopedia of the Social Sciences (1968)

17. Croux, C., Dehon, C.: Influence functions of the spearman and kendall correlation measures. Statistical methods & applications 19(4), 497–515 (2010)

18. for Ethical AI, T.I., Learning, M.: Ethical XAI (2021), https://ethical.institute/xai.html

19. Främling, K.: Decision theory meets explainable ai. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 57–74. Springer (2020)

20. Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature 521(7553), 452–459 (2015)

21. GILABERTE, B.: Crimes contra a pessoa. Rio de Janeiro: Freitas Bastos (2013)
22. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). pp. 80–89 (2018). https://doi.org/10.1109/DSAA.2018.00018

23. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24(1), 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095, https://doi.org/10.1080/10618600.2014.907095

24. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5), 1–42 (2018)

25. Gunning, D., Aha, D.: DARPA’s Explainable Artificial Intelligence (XAI) Program. AI Magazine 40(2), 44–58 (Jun 2019). https://doi.org/10.1609/aimag.v40i2.2850

26. Gunning, D., Stefk, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: Xai—explainable artificial intelligence. Science Robotics 4(37), eaay7120 (2019)

27. Hall, P., Gill, N., Kurka, M., Phan, W.: Machine learning interpretability with h2o driverless ai. H2O. ai (2017)

28. Hong, S.R., Hullman, J., Bertini, E.: Human factors in model interpretability: Industry practices, challenges, and needs. Proceedings of the ACM on Human-Computer Interaction 4(CSCW1), 1–26 (2020)

29. IBGE: https://cidades.ibge.gov.br/brasil/pa/belem/panorama, Last accessed 04 Mar 2021.

30. IPEA: www.ipea.gov.br/atlasviolencia/arquivos/downloads/7047-190802atlasdaviolencia2019municipios.pdf, Last accessed 04 Mar 2021.

31. Jech, T.J., Jech, T., Jech, T.J., Mathematician, G.B., Jech, T.J., Mathématicien, G.B.: Set theory, vol. 14. Springer (2003)

32. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., Wortman Vaughan, J.: Interpreting interpretability: Understanding data scientists’ use of interpretability tools for machine learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–14 (2020)

33. Keeney, R.L., L, K.R., Howard, R.: Decisions with Multiple Objectives: Preferences and Value Trade-Offs. Cambridge University Press, Cambridge England ; New York, NY, USA, revised ed. edição edn. (Aug 1993)

34. Kiran, M., Khan, S.: Machine learning for safety–critical applications in engineering. Machine Learning 109(5), 1101–1102 (2020)

35. Korobov, M., Lopuhiin, K.: Eli5. https://eli5.readthedocs.io/en/latest/index.html, Last accessed 21 Jan 2021.

36. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy 23(1) (2021). https://doi.org/10.3390/e23010018 https://www.mdpi.com/1099-4300/23/1/18

37. Lipovetsky, S., Conklin, M.: Analysis of regression in game theory approach. Applied Stochastic Models in Business and Industry 17(4), 319–330 (2001)

38. Loukas, S., Papaioannou, T.: Rank correlation inequalities with ties and missing data. Statistics & probability letters 11(1), 53–56 (1991)

39. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence
40. Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence 2(1), 2522–5839 (2020)
41. Messalas, A., Kanellopoulos, Y., Makris, C.: Model-agnostic interpretability with shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA). pp. 1–7 (2019).
42. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
43. Nguyen, T.N., Choo, R.: Human-in-the-loop xai-enabled vulnerability detection, investigation, and mitigation. In: 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE). pp. 1210–1212. IEEE (2021)
44. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
45. da Paz, I.S.: Onde mora a impunidade?
46. Reza, F.M.: An Introduction to Information Theory. Courier Corporation (Jan 1994), google-Books-ID: RtzpRAiX6OgC
47. Ribeiro, J., Meneses, L., Costa, D., Miranda, W., Alves, R.: Prediction of homicides in urban centers: A machine learning approach. In: Arai, K. (ed.) Intelligent Systems and Applications. pp. 344–361. Springer International Publishing, Cham (2022)
48. Ribeiro, J., Silva, R., Cardoso, L., Alves, R.: Does dataset complexity matters for model explainers? In: 2021 IEEE International Conference on Big Data (Big Data). pp. 5257–5265 (2021). https://doi.org/10.1109/BigData52589.2021.9671630
49. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA. August 13-17, 2016. pp. 1135–1144 (2016)
50. Roth, A.E.: The Shapley value: essays in honor of Lloyd S. Shapley. Cambridge University Press (1988)
51. Shalev-Shwartz, S., Ben-David, S.: Understanding machine learning: From theory to algorithms. Cambridge university press (2014)
52. Silva, A.: Projeto e analise de algoritmos. Instituto Militar de Engenharia, Rio de Janeiro (1998)
53. Singh, J., Anand, A.: Model agnostic interpretability of rankers via intent modelling. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 618–628 (2020)
54. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 56–67 (2020)
55. Sokol, K., Flach, P.: Explainability fact sheets: a framework for systematic assessment of explainable approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 56–67 (2020)
56. Sokol, K., Flach, P.: One explanation does not fit all. KI-Künstliche Intelligenz 34(2), 235–250 (2020)
57. Stevens, S.S., et al.: On the theory of scales of measurement (1946)
58. Wiens, J., Wallace, B.C.: special issue on machine learning for health and medicine. Machine Learning 102(3), 305–307 (2016)
59. Wilming, R., Budding, C., Müller, K.R., Haufe, S.: Scrutinizing xai using linear ground-truth data with suppressor variables. Machine learning pp. 1–21 (2022)