Evaluation Metrics in Explainable Artificial Intelligence (XAI)

Chapter · November 2022

DOI: 10.1007/978-3-031-20319-0_30

2 authors:

Loredana Coroamă
Universitatea Tehnica Cluj-Napoca
7 PUBLICATIONS **7** CITATIONS
SEE PROFILE

Adrian Groza
Universitatea Tehnica Cluj-Napoca
185 PUBLICATIONS **438** CITATIONS
SEE PROFILE

Evaluation metrics for Explainable Artificial Intelligence techniques: State of the Art Review and Challenges

Loredana Coroama[1] and Adrian Groza[1]

***Abstract*—Although AI is spread across all domains and many authors stated that providing explanations is crucial, another question comes into play: How accurate are those explanations? This paper aims to summarize a state-of-the-art review in XAI evaluation metrics, to present a categorization of evaluation methods and show a mapping between existing tools and theoretically defined metrics by underlining the challenges and future development. The contribution of this paper is to help researchers to identify and apply evaluation metrics when developing an XAI system and also to identify opportunities for proposing other evaluation metrics for XAI.**

***Keywords*-explainable artificial intelligence, interpretability, explanation methods, explanation quality, explanation metrics**

I. INTRODUCTION

Artificial intelligence has experienced a significant growth in the last decade, especially in critical decision-making systems (e.g. medicine or justice related applications), recommendation systems or different processes such as credit lending or employment. An issue arises when the system rely upon black box models such as Machine Learning and Deep Neural Networks rather than a simple statistical model because increasing the model complexity comes with the trade-off of decreasing the interpretability, therefore, quantifying and explaining the effectiveness of XAI to the people that system interacts with becomes a challenge.

There is also an increasing demand for a responsible and accountable AI that are achieved by providing explainability. Explainable AI systems aims to detect unwanted biases such as gender or social discrimination ensuring fairness, safety, privacy through algorithmic transparency. The main issue is that most algorithms collect and analyze user data affecting decision making. Each individual should have a right to an explanation according to GDPR commission. One negative recent example was in political elections where personal information of people has been collected and used against their will. Their feed was flooded with spam during political elections to influence their decision. The strategy of politicized social bots is outlined by Samuel C. Woolley in his study [1].

Another great advantage of providing explanations is gaining user trust. A system that provides explanations is perceived more human-like by users because it is part of human nature to assign causal attribution of events. Figure 1 shows the architecture of XAI process where for each decision generated by an intelligent system explanations are provided to the user increasing trust and confidence. It is very important to asses the quality of explanations as many evaluation metrics are addressed in the literature.

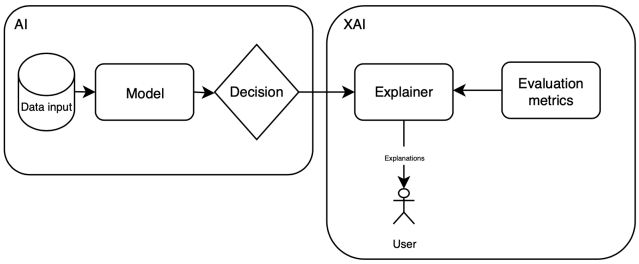

Figure 1: XAI process

XAI is applied in various industrial applications. In recommendation systems, recommended posts and news are explained to establish trust between users and content providers. In sales, prediction are made to upsell or churn and the most relevant features are perturbed in order to influence the result. In human resources, it might be needed to discriminate between many candidates for a certain role. In medicine, one example is the Predict tool developed by David Spiegelhalter [2] that shows how different treatments for breast cancer might improve survival rates after surgery. In credit lending, counterfactual explanations are mandatory. Another usages are in explaining energy consumption or critical systems such as object detection.

The paper is organised as follows. Classification of XAI metrics, methods and existing studies are presented in Section II, existing XAI evaluation metrics are discussed in Section III, while implementations are describes in Section IV and conclusions are drawn in Section V.

II. XAI CLASSIFICATIONS

Regarding the state of the XAI research, existing studies are split in three main categories. Some papers review existing XAI methods or propose new techniques for explainability ([3], [4], [5]). Second category is focused on the notions of explainability ([6], [7], [8]), while the third

[1]Computer Science Department Technical University of Cluj-Napoca, 28 Memorandumului, Cluj-Napoca, Romania adrian.groza@cs.utcluj.ro, loredana.coroama@campus.utcluj.ro

one tends to evaluate all these approaches proposing new evaluation metrics([9], [10], [11], [12], [13]).

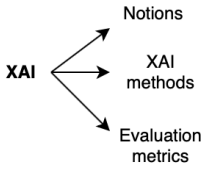

Figure 2: Current state of XAI research

Giulia Vilone and Luca Longo [7] classify existing XAI methods based on five criterias. The first one is related to the stage in which explanations are generated. Explanations can be generated during training or a post-model (agnostic or specific) can be built which is less computationally expensive. The second one deals with the type of explanations which can be global or local. The third one split the methods based on the problem type (classification and regression). Also, input data is important because not all the methods works for all data types such as text, images or tabular data. Finaly, one more criteria is related to the output format. Explanations can be numerical, textual, visual, mixted or rules-based. They also performed a classification of evaluation metrics. According to the authors, there are human-centred and objective metrics. Human-centred metrics involve users feedback. Users can be randomly selected or domain experts. Also, depending on the questions addressed to people, there are qualitative metrics which aim to achieve deeper insights and quantitative metrics for statistical analysis. On the other hand, objective metrics such as explanation completeness, rules cardinality or perturbation metrics are defined based on formal definitions. Explanation completeness captures the highest number of features that determine the prediction while perturbation metrics refers to the sensitivity to input perturbation by altering the input and comparing the outputs and to model parameter randomization by comparing with same models but with different parameters.

Finale Doshi-Velez and Been Kim [8] divided XAI evaluation in tree categories: user-based evaluation, application-based evaluation and functionality-based. They also considered first two types as a part of human-centred evaluation and split it in subjective and objectives metrics. Moreover, we claim that functionality-based evaluation which assesses the quality of explanations using formal definitions of explainability could be divided in method-specific metrics and agnostic metrics.

Other evaluation approach is related to features perturbation. Perturbing relevant features should lead to a change in prediction, therefore higher the change, better the method. Secondly, example-based explanations seems to be another type of evaluation that is better designed for humans ([14], [5]). Another category of explanations depends on data, Peter Hase and Mohit Bansal [15] performed experiments by generating forward and counterfactual explanations. On the other side, [16] shows that visual explanations are not always of use. Also, explanations can be evaluated from a developer point of view [17].

Taking into consideration related work discussed in Section III, we categorize existing XAI evaluation metrics as it is shown in Figure 3. We consider the metrics that rely on human feedback as subjective. The feedback can be provided by randomly selected persons or by domain experts such as doctors in medicine or judges in justice. They are subjective as each user might consider that a particular method or application is more appropriate for them relying also on different aspects including user experience or application design which does not necessarily assess the explanation quality. On the other hand, objectivity could be achieved through formal definitions. Many explanations methods have been developed in the past years, some metrics were developed only for specific methods or tasks, while others are focused on the model behaviour. Example-based methods provide explanations by generating the most similar instances for the sample being explained, therefore aspects such as non-representativeness or diversity stand as valid options for quantifying the effectiveness of explanations. Counterfactual explanations are used when knowing the outcome is not enough and explanations are required to determine what should be changed for obtaining another result. These could be evaluated with metrics such as diversity of changes and their feasibility. Other specific metrics could be related to a certain task such as generating explanations in recommendation systems. The last category covers model-related metrics. Attribution-based metrics are computed taking into consideration feature importance. Examples of such metrics are sensitivity or the monotony. Some methods build another oversimplificated version of the initial model (post-model) in order to generate explanations which is efficient because avoids model retraining and reduces the computational cost. In this case, evaluating the quality of explanations comes down to the post-model evaluation with metrics such as size, complexity or accuracy. Performance aspects refers to metrics such as computational cost which can be highly expensive or even stopping the execution of the algorithm for very large datasets. Model trustworthiness is another aspect that comes into play. Metrics such as consistency or stability increase confidence when quantifying the quality of explanations.

Numerous metrics underlying the quality of explanations were proposed in the literature as summarised in Table I.

III. Discussion and Related work

If an XAI agent works in tandem with a specialist then a maximum accuracy can be achieved. Avi Rosenfeld considers state-of-the-art imaging techniques in radiology in order to identify diseases [12]. Radiologists have a successful rate of 97 % in finding the disease, while the agent has achieved

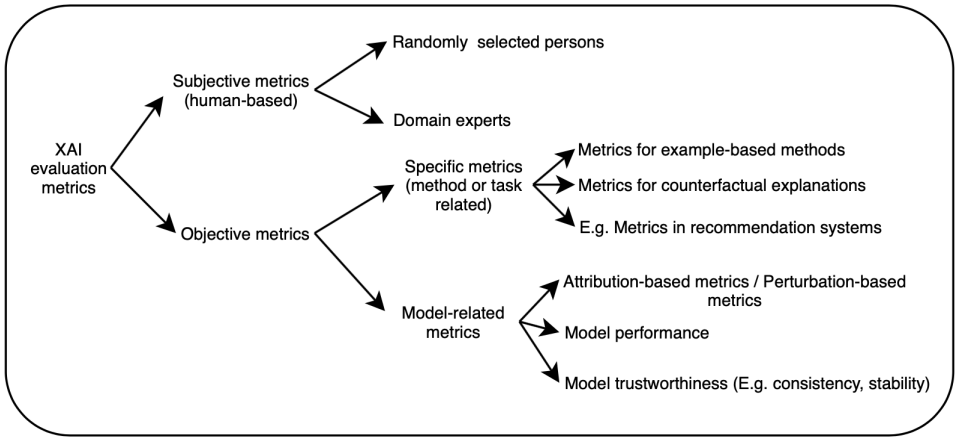

Figure 3: XAI evaluation metrics

Table I: Existing metrics

XAI metric	Type	Observations	Implementation
D	Objective	The performance difference between the agent's model and the logic presented as an explanation	No
R	Objective	The number of rules in the agent's explanation, the fewer, the better	No
F	Objective	The number of features used to construct the explanation, less features, clearer the explanation	No
S	Objective	The stability of the agent's explanation	Yes
Simplicity	Objective	The ability to choose only the necessary and sufficient features for explaining the prediction	No
Sensitivity	Objective	Measure the degree to which the explanation is affected by insignificant perturbations from the test point	Yes
Completeness	Objective	Captures the highest number of features that determine the prediction	No
Soundness	Objective	How truthful each element in an explanation is	No
Stability of explanation	Objective	Test the consistency of explanation methods consisting in many repeated experiments; Close inputs with similar predictions yields similar explanations	Yes
Robustness	Objective	Similar inputs should result in similar explanations, measure the sensitivity to noise	Yes
Computational cost	Objective	How expensive is to generate explanations	Yes
Post-model evaluation metrics	Objective	Model size (number of rules, length of rules or depth of trees), model complexity, interaction strength, etc.	Yes
Monotonicity	Objective	It is measured by adding each feature in order of increasing importance and observing an increase in the model performance; Features attributions should be monotonic, otherwise the correct importance of the features is not provided	Yes
Perturbation-based metrics	Objective	The capacity of underlying the variations after perturbing the inputs	Yes
Non-representativeness	Objective	Measure of the fidelity of the explanation; used in example-based methods	No
Diversity	Objective	Used in example-based methods	No
Explanation correctness	Objective	Assessed through sensitivity and fidelity	Yes
Explanation confidence	Objective	It is concerned with whether the generated explanation and the masked input result in high confidence predictions.	Yes
Fidelity (MuFidelity, Deletion and Insertion)	Objective	Explanations describe correctly the model behaviour; Ensure there is a correlation between a random subset of pixels and their attribution score	Yes
Representativity (Generalizability)	Objective	How much seeing one explanation informs you about the others; the more representative an explanation is, the more it persists when we remove a point	Yes
Consistency	Objective	The consistency score of the explanations, it informs about the confidence of the explanations, how much two models explanations will not contradict each other; The explainer should capture the same relevant components under various transformation to the input	Yes
Faithfulness	Objective	Computes which feature has the most impact on the model output when individually changed; Removing each important feature should result in decreasing the model performance	Yes
ROAR	Objective	Retrains the model with the most relevant features removed	Yes
GT-Shapley	Objective	Determine which technique compute most accurately the approximations to the Shapley values	Yes
Infidelity	Objective	The difference between the change in function value and the dot product of the change in feature value with the feature importance vector, considering that each feature is replaced with a noisy baseline conditional expectation	Yes
Metrics for counterfactual explanations (diversity, feasibility, validity, sparsity)	Objective	Diversity implies a wide range of suggested changes and feasibility the possibility to adopt those changes; Validity measures the uniqueness and the sparsity refers to the number of features that are different	Yes
Transparency	Subjective	Describe how the system takes a certain decision; Used in recommendation systems (not only)	No
Scrutability (similar with actionability or correctability)	N/A	Ability to to correct the system if its assumptions are wrong; Used in recommendation systems	No
Trust	Subjective	Measured through user questionnaires or metrics such as products sold in recommendation systems	No
Effectiveness	Subjective	Used in recommendation systems to discard unwanted options	No
Persuasiveness	Subjective	Used in recommendation systems to convince the user to take an action (e.g. buy a product) after receiving the explanations	No
Efficiency	Subjective	Used in recommendation systems / conversational systems and can be measured by counting the explanations needed	No
Satisfaction	Subjective	Usefulness and ease of use; used in recommendation systems	No
Comprehensibility	Subjective	How much effort is required for humans to understand the explanations	No
Justifiability	Subjective	Assess if the model is in line with domain knowledge	No
Explanation goodness, User curiosity/attention engagement, User understanding, User performance/productivity, System controllability/interaction, Explanation usefulness	Subjective	Used in psychology	No
Interactivity, interestingness, informativeness, human-AI task performance	Subjective	User experience	No

99.5 % accuracy. As this area is very sensitive and the decisions are crucial, it would be ideal to have meaningful explanations for those misclassified samples; in this manner, doctors with knowledge expertise will identify the mistakes. In presented paper, four metrics are suggested in order to evaluate generated explanations: i) D, the performance difference between the agent's model and the performance of the logic presented as an explanation. ii) R, the number of rules in the agent's explanation. iii) F, the number of features

used to construct the explanation. iv) S, the stability of the agent's explanation. The author claims that these objective metrics are critical raising the issue that existing studies using methods such as post-models are a oversimplication of the initial models logic and they didn't evaluate legal, ethical or safety concerns. Also, an advantage of the metrics developed is that they are not dependant of the task being performed or the XAI algorithm.

Black-box algorithms such as neural networks achieve a higher accuracy than white box algorithms, but are less transparent. D measures if choosing black box algorithms is really necessary. It quantifies the change of agent performance (δ) between the black box model (P_b) and the transparent model (P_t). The decision is made comparing $P_b - \delta$ and P_t. R focuses on the size of the explanations, the fewer, the better: $\lambda * L$, where $L = [size(m) - c]$, m is number of rules and c a penalization. This metric is more suitable for transparent methods. F focuses on the number of features used to create the explanation, less features, clearer the explanation. This metric is quantified similarly with the previous one except that represents the number of inputted features instead of outputted rules. S is the stability of the explanations which quantifies the ability to handle small noise perturbations: $\lambda * (1 - similarity)$. Some metrics for quantifying similarity are Jaccard and Tanimoto.

There are three ways to generate explanations: directly using a transparent algorithm, through feature selection or by creating a post-model. The last option aims to describe the inner working of a black-box algorithm that is not inherently understood. This is achieved by approximating model's logic via white-box algorithms or by highlighting superpixels or using model perturbations. Also, some authors consider a difference between explainability and interpretability claiming that explainability focuses on the ability of humans to understand model logic and interpretability clarifies the system's internal logic. On the other side, in many papers these terms are used synonymously.

In recommendation systems, Nava Tintarev et al. [4] claim that metrics such as user satisfaction, serendipity, diversity and trust are as important as accuracy metrics such as precision and recall. To asses explanation quality, they have proposed the following metrics: transparency, scrutability, trust, effectiveness, persuasiveness, efficiency and satisfaction. These metrics were analysed in many recommendation systems. Transparency should explain how a system works and why it recommended a certain decision. There is no work on evaluating transparency. Scrutability allows users to correct the system if its assumptions were wrong. It is also known as user control. It was proved that some recommendations are made based on the user profile, therefore changing his attributes would result in generating different recommendations. Trust is linked with transparency and accuracy, but sometimes explanations can compensate faulty recommendations. Moreover, Bj Fogg et al. [18] demonstrated that the design of the system might affect the users credibility. Trust has been measured through user questionnaires or user loyalty by counting metrics such as the number of logins or products sold. Persuasiveness is a technique to convince the user to buy or try and it can be measured after the users receive explanations. Effectiveness help users to discard unwanted options by generating good explanations. Efficiency is mostly used in conversational systems and measures how quickly a task can be performed. An evaluation metric could be the number of explanations needed. Satisfaction is related to usefulness and ease of use and can be evaluated using user-based metrics.

Giulia Vilone and Luca Longo [7] have presented many notions related to the explanation quality. Actionability and correctability describe the capacity to transfer knowledge to the users allowing them to explain necessary corrections. Causality [19] refers to the relationship between input and output. Completeness defines the extent to which a system is described by explanations. Comprehensibility [20] quantifies how much effort is required for humans to understand the explanations. Faithfulness represents the capacity to select truly relevant features. Justifiability assess if the model is in line with domain knowledge. Robustness means that similar inputs should result in similar explanations, David et al. [21] quantifies this metric and claim that current XAI methods do not perform well. Scrutability allows inpecting a training process that fails. Simplicity is the ability to choose only the necessary and sufficient features for explaining the prediction. Sensitivity reflects the capacity of underlying the variations after perturbing the inputs. Stability measures the consistency. Soundness analyses how truthful each element in an explanation is. Todd Kulesza et al. [22] claim that completeness is more important than soundness. Many other notions and metrics such as transparency, effectiveness, efficiency, interactivity, interestingness, informativeness, persuasiveness, satisfaction and security are discussed, but there are no formal definitions or implementations developed.

Other domains in which XAI research takes advantages are cognitive science and psychology. 7 cognitives metrics were introduced by Janet Hui-wen Hsiao et al. in their study [13]: explanation goodness, user satisfaction, user curiosity/attention engagement, user trust/reliance, user understanding, user performance/productivity and system controllability/interaction. They claim that existing evaluation methods are inherited from cognitive processes. There are 3 processes in cognitive science and psychology: measuring and comparing the system's behavior under different conditions (e.g. perturbation-based methods), building predictive models to simulate the behaviour (XAI post-models) and explaining the behaviour by analysing features through factor analysis (XAI methods based on feature importance). Presented metrics can be measured using subjective methods as they require user interaction.

Sina Mohseni et al. [10] classified evaluation methods

based on targeted users such as AI novices, domain experts and AI experts. Primary interpretability measures refers to user's mental model, explanation usefulness and satisfaction, user trust and reliance and the human-AI task performance. Last type covers the computational measures such as explanation correctness which is strong related to model consistency, explainer fidelity and model trustworthiness and do not rely on human-subject studies.

Stability of explanation, robustness of referee classifier and computational cost are discussed when quantifying the informativeness of explanation methods for time series classification [23]. T. Nguyen et al. studied the saliency maps method for explanations which highlight the most important parts when making a prediction. They define an explanation as being truly informative if it points out the part of time series that are most relevant for the prediction. The stability of explanation test the consistency of explanation methods consisting in many repeated experiments. Robustness measures how sensitive to noise are presented methods. The computational costs measures how expensive is to generate explanations.

Jianlong Zhou et al. [9] define explainability as a combination of interpretability (explanations understandable by humans) and fidelity (explanations describe correctly the model behaviour). They claim that general computation metrics for XAI methods evaluation is unlikely to be possible due to many factors such as the subjectivity of explanations, the context, the models, the users dependency and the type of explanations required. They also divided objective evaluation metrics in tree types: model-based explanations, attribution-based explanations and example-based explanations. Model-based explanations use the model itself or create new models to explain ML. Evaluating the quality of explanations means evaluating that model. Examples of such metrics are model size (number of rules, length of rules or depth of trees), interaction strength or model complexity. Attribution-based explanations are based on feature importance or feature ranking. Examples of metrics are monotonicity or sensitivity. Example-based explanations select most similar instances from the dataset. Examples of metrics are non-representativeness and diversity [24].

Alejandro B. Arrieta et al. [6] suggest as future improvement the development of certain evaluation metrics: the goodness, usefulness and satisfaction of explanations, the improvement of the mental model of the audience induced by model explanations and the impact of explanations on the performance of the model and on the trust and reliance of the audience. Goodness checklist, explanation satisfaction scale, elicitation methods for mental models, computational measures for explainer fidelity, explanation trustworthiness and model reliability are described in [11], [25].

The reasoning of Bayesian Networks has been applied in different applications. BayLime [26] is an extension of LIME, a well-known XAI technique that it was proved to be unstable do to the lack of consistency when generating different explanations for same instances. BayLime takes advantage of prior knowledge and Bayesian reasoning to improve the consistency and the robustness to kernel settings. The need for explaining bayesian networks for legal evidence was also addressed [27]. Carmen Lacave et al. [3] review the explanation methods for Bayesian networks. They defined some properties that each explanation should satisfy. There are three aspects that should be taken into consideration: the content (what to explain), communication (how the system interacts with the user), adaptation (to whom the explanation is addressed). The content covers the focus, purpose, level and causality of explanations. First property is the focus of explanation and defines 3 basic issues that should be explained by any expert system: the knowledge base (explaining the model), the reasoning (explaining obtained results and the reasoning process or hypothetical reasoning) and the evidence (which unobserved variables justify the available evidence). The purpose refers either to the description of the model or other aspects or the comprehension which explain how each finding affects the results. The level of explanation in Bayesian networks consists in micro-level explanations by generating detailed explanations for a particular node or macro-level explanations which analyses the main lines in the network. Causality is one of the most important features because humans tend to interpret events in terms of cause-effects relations. The communication represents the way in which it is offered to the users and covers the user-system interaction, display of explanations (text, numbers or graphs) and expressions of probability (numerical, quantitative or linguistic such as "very likely"). The adaptation consists in the ability to address each user's needs depending on the knowledge he has and covers the user's knowledge about the domain and reasoning method and the level of detail. There are various tools for explainining Bayesian Networks (Hugin [28], Analytica, Ideal [29], David [30], Diaval [31], Elvira[1], Medicus [32], B2 [33], Banter [34]), but no metrics developed to assess their correctability.

IV. Implementations

Predict Tool[2] [2] helps doctors and patients to decide which treatment to take after breast cancer surgery. The tool explains its decisions and provides graphics, text and tables with how that treatment affected other women after surgery. Similar tools could be developed for other types of cancer.

There are different ways to explain a decision. Different users require different explanations for different purposes and with different objectives. A doctor wants to know why a certain treatment is recommended (trust), the government wants to prove that there is no discrimination (compliance,

[1]Elvira tool available online: https://leo.ugr.es/elvira/

[2]PredictTool available online: https://breast.predict.nhs.uk/tool

safety) and a developer is concerned about how is the system performing or how it can be improved (quality, debug). In all the cases, the explanations quality is crucial. IBM AIX360 [35] developed a similar application[3] with 3 types of users: developer (ensure the model works appropriately), loan officer (needs to assess the model's prediction and make the final judgement), bank customer (wants to understand the reason for the application result). They also implemented 2 metrics in order to evaluate the explanations methods: faithfulness and monotony. Faithfulness measures the correlation between the model attributes and the model performance. Removing each important feature should result in decreasing the model performance. On the other hand, monotony is measured by adding each feature in order of increasing importance and observing an increase in the model performance.

Main drawbacks when developing new metrics are: computational cost, inability to be extended to non-image domains or simply focusing only one desirable attribute of a good explainer.

XAI-Bench tool [36] uses synthetic datasets to evaluate the faithfulness, monotonicity, ROAR, GT-Shapley and infidelity of five XAI techniques. Faithfulness computes which feature has the most impact on the model output when individually changed, while monotonicity computes the effect of the features added sequentially. Remove-and-retrain (ROAR) consists in retraining the model with the most relevant features removed, while GT-Shapley metric determine which technique compute most accurately the approximations to the Shapley values. Infidelity computes the difference between the change in function value and the dot product of the change in feature value with the feature importance vector, considering that each feature is replaced with a noisy baseline conditional expectation.

DiCE [37] is a tool that generates and evaluates counterfactuals explanations. In some cases such as credit lending, knowing the outcome is not enough, the applicant might want to know what to do to obtain a better outcome in the future. The metrics developed are diversity and feasibility. Diversity implies a wide range of suggested changes and feasibility the possibility to adopt those changes (proximity to the original input). Other constraints are validity (unique examples which correspond to a different outcome than the original input) and sparsity (the number of features that are different).

Implementations for correctness, consistency and confidence [4] are developed in the literature. For example, consistency refers to the fact that the explainer should capture the same relevant components under various transformation to the input. Correctness is the same thing with sensitivity and fidelity and represents the ability of the explainer to determine the most relevant features. Confidence is concerned with whether the generated explanation and the masked input result in high confidence predictions. This metrics are computationally inexpensive because do not require model retraining.

Xplique[5] is a tool dedicted to explainability for neural networks based on Tensorflow. It contains a module that allow testing the current evaluation metrics. Fidelity is addressed through 3 metrics: MuFidelity, Deletion and Insertion. The deletion metric measures the drop in the probability of a class as the input is gradually perturbed. The insertion metric captures the importance of the pixels in terms of their ability to synthesize an image and is measured by the rise in the probability of the class of interest as pixels are added according to the generated importance map. MuFidelity ensure there is a correlation between a random subset of pixels and their attribution score. Stability ensure that close inputs with similar predictions yields similar explanations. Representativity (or Generalizability) gives you an overview of the generalization of your explanations: how much seeing one explanation informs you about the others. Consistency is the consistency score of the explanations, it informs about the confidence of the explanations, how much two models explanations will not contradict each other. Explicitness, faithfulness, stability are also presented in [38].

V. Conclusions and future improvements

The integration and evaluation of XAI methods has become very important as the AI is more and more widespread in almost all domains. In the last years, artificial intelligence has been used in unethical purposes such as computational propaganda, fake news spreading and different campaign to manipulate public opinion. Two well-known events are the war in Ukraine and COVID-19 in China. Some datasets are provided by the university of Oxford[6] which can be used for future research.

This paper summarizes existing evaluation methods and metrics in the literature underlining two issues. A big part of the research limits itself to theoretical definitions that have not been implemented or adopted in real applications. Then it seems that general evaluation metrics for XAI methods are unlikely to be implemented due to many factors such as the task being performed, the internal logic of the XAI method or the type of the explanations. However, some metrics could be generally applicable (e.g. each method should be consistent and return same explanations at different runs for the same sample).

Moreover, many metrics come with a trade-off. It is desired to achieve all these features at their maximum capacity, but in reality it might be impossible (e.g. an increase in transparency comes with the decrease in efficiency and so

[3]IBM AIX360 available online: http://aix360.mybluemix.net/data

[4]https://github.com/amarogayo/xai-metrics

[5]https://github.com/deel-ai/xplique

[6]https://demtech.oii.ox.ac.uk/research/data-sets/

on). The evaluation metrics should be mapped to the system goal and what the explanation is trying to achieve.

REFERENCES

[1] Samuel C. Woolley. Automating power: Social bot interference in global politics. *First Monday*, 21(4), Mar. 2016.

[2] David Spiegelhalter. Making algorithms trustworthy: What can statistical science contribute to transparency, explanation and validation? *NeurIPS*, 2018.

[3] Carmen Lacave and Francisco Dez. A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17, 05 2001.

[4] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. pages 801–810, 05 2007.

[5] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *34th Conference on Neural Information Processing Systems*, 2020.

[6] 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. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[7] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review, 05 2020.

[8] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.

[9] Jianlong Zhou, Amir H Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

[10] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *CoRR*, abs/1811.11839, 2018.

[11] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[12] Avi Rosenfeld. Better metrics for evaluating explainable artificial intelligence: Blue sky ideas track. 05 2021.

[13] Janet Hui-wen Hsiao, Hilary Hei Ting Ngai, Luyu Qiu, Yi Yang, and Caleb Chen Cao. Roadmap of designing cognitive metrics for explainable artificial intelligence (XAI). *CoRR*, abs/2108.01737, 2021.

[14] Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain CNN activations better than feature visualizations. *CoRR*, abs/2010.12606, 2020.

[15] Peter Hase and Mohit Bansal. Evaluating explainable AI: which algorithmic explanations help users predict model behavior? *CoRR*, abs/2005.01831, 2020.

[16] Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? A case study in model-in-the-loop prediction. *CoRR*, abs/2007.12248, 2020.

[17] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. *CoRR*, abs/1909.06342, 2019.

[18] Bj Fogg, Jonathan Marshall, Tami Kameda, Joshua Solomon, Akshay Rangnekar, John Boyd, and Bonny Brown. Web credibility research: A method for online experiments and early study results. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, pages 295–296, 2001.

[19] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery*, 9, 2019.

[20] I. Askira-Gelman. Knowledge discovery: Comprehensibility of the results. In *2014 47th Hawaii International Conference on System Sciences*, volume 5, page 247, Los Alamitos, CA, USA, jan 1998. IEEE Computer Society.

[21] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.

[22] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, 2013.

[23] Thu Trang Nguyen, Thach Le Nguyen, and Georgiana Ifrim. A model-agnostic approach to quantifying the informativeness of explanation methods for time series classification. In *International Workshop on Advanced Analytics and Learning on Temporal Data*, pages 77–94. Springer, 2020.

[24] An-phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. *CoRR*, abs/2007.07584, 2020.

[25] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A multidisciplinary survey and framework for design and evaluation of explainable ai systems, 2020.

[26] Xingyu Zhao, Xiaowei Huang, Valentin Robu, and David Flynn. Baylime: Bayesian local interpretable model-agnostic explanations. *CoRR*, abs/2012.03058, 2020.

[27] Charlotte S Vlek, Henry Prakken, Silja Renooij, and Bart Verheij. A method for explaining bayesian networks for legal evidence with scenarios. *Artificial intelligence and law*, 24(3):285–324, 2016.

[28] Stig Andersen, Kristian Olesen, Finn Jensen, and Frank Jensen. Hugin - a shell for building bayesian belief universes for expert systems. volume 2, pages 1080–1085, 01 1989.

[29] Sampath Srinivas and John S. Breese. IDEAL: A software package for analysis of influence diagrams. *CoRR*, abs/1304.1107, 2013.

[30] Ross D. Shachter. DAVID: influence diagram processing system for the macintosh. *CoRR*, abs/1304.3108, 2013.

[31] Francisco Díez, J Mira, E Iturralde, and S Zubillaga. Diaval, a bayesian expert system for echocardiography. *Artificial intelligence in medicine*, 10:59–73, 06 1997.

[32] Olaf Schröder, Claus Möbus, and Heinz-Jürgen Thole. Knowledge from linguistic models in complex, probabilistic domains. 01 1996.

[33] Susan Mcroy, Alfredo Liu-perez, and Susan Haller. B2: A tutoring shell for bayesian networks that supports natural language interaction. 02 1996.

[34] Peter Haddawy, Joel Jacobson, and Charles E Kahn. Banter: a bayesian network tutoring shell. *Artificial Intelligence in Medicine*, 10(2):177–200, 1997.

[35] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, 2019.

[36] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Advances in Neural Information Processing Systems Datasets Track*, 2021.

[37] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[38] David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. *CoRR*, abs/1806.07538, 2018.