Literature Review on the State of Explainable AI in Medicine

Karman Singh

Introduction

Artificial Intelligence and Machine Learning have demonstrated remarkable potential in numerous domains from beating humans at Go to self-driving cars. Most of the recent growth in machine learning has been driven by the widespread use of complex models, like deep neural networks. However, this complexity comes at a cost as these ML systems are black boxes. Users and people being impacted have little to no understanding of how they make predictions. This lack of understanding presents multiple problems with serious consequences, and we need to develop ethical models that are interpretable, tractable, and trustworthy.

The use of ML systems is expanding not just in software engineering but into education, law enforcement, and healthcare. In health care, current AI-based systems fail to hold up performance in the real-world clinical environments. Other concerns surrounding the use of ML in medicine include privacy, bias, lack of transparency, and security as well as causality, informativeness, and fairness. AI decisions made or influenced by such systems affect human health, there is an urgent need for understanding of how such decisions are made.

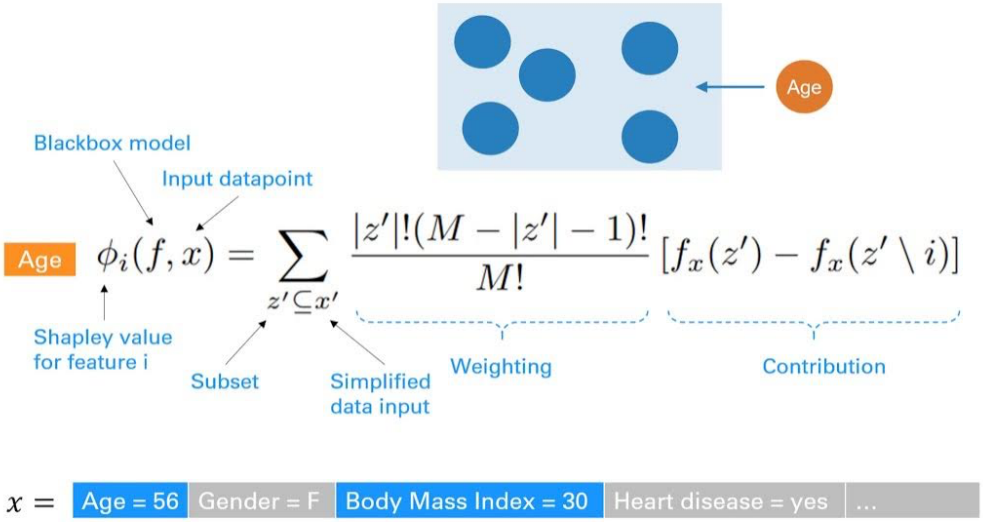
A consensus is emerging in favor of explainable AI/ML to highlight how predictions are made, but we need to be wary about explainable AI in its current. In this report, I will highlight four papers that touch on an overview of limits in explainability, the definitions of explainability and how to evaluate them, the importance of extending explainability to causability, an example of explainable AI evaluation on a multi-modal medical imaging task, and how explanation trustworthiness is overlooked.

What is Explainable AI?

Explainable AI/ML, as the term is typically used, does roughly the following: Given a black-box model that is used to make predictions, a second explanatory algorithm finds an interpretable function that closely approximates the outputs of the black box. This second algorithm is trained by fitting the predictions of the black box and not the original data and is typically used to develop the post hoc explanations for the black-box outputs. The explanation might, for instance, be given in terms of which attributes of the input data in the black-box algorithm matter most to a specific prediction. In similar terms, explainable AI finds a white box that partially mimics the behavior of the black box, which is then used as an explanation of the black-box predictions.

Two of the most popular explanation methods uses are LIME and SHAP. LIME stands for Local Interpretable Model-agnostic Explanations. Model agnosticism refers to the property that LIME can give explanations for any given supervised learning model. Local explanations mean that LIME gives explanations that are locally faithful within the surroundings of sample being explained. SHAP

stands for Shapley Additive Explanations, which as method to explain individual predictions, based on the game theoretically optimal Shapley values. Shapley values are a widely used approach from cooperative game theory that come with desirable properties. The feature values of a data instance act as players in a coalition. The Shapley value is the average marginal contribution of a feature value across all possible coalitions. The figure below provides more details on how the Shapley value is calculated.



Limits of Explainability

Explainable AI outputs post hoc rationalizations of a black box predictions, which aren't necessarily the actual reasons behind the prediction. This means they are unlikely to contribute to our understanding of its inner workings. Instead, we are left with the false impression that we understand them better. We call the understanding that comes from post hoc rationalizations "ersatz understanding." This type of understanding won't improve trust or diminish any underlying moral, ethical, or legal concerns. Another issue is the Lack of robustness in explanations. For a small change in input, like a pixel in an image, an explainable method might produce very different explanation. A doctor using AI-based medical device would naturally question that algorithm. The explanations might be misleading in the hands of imperfect users as the average user is vulnerable to narrative fallacies, where users combine and reframe explanations. Also, simple post hoc rationale can engender a false sense of overconfidence as it can be difficult to understand the full rationale. Finally, when we use more interpretable white box models to favor more explainability, we are often trading accuracy.

Definitions of Explainability

This next paper goes into expanding and defining explainability and how to evaluate explanations. Explainability is subjective and the perceived quality of an explanation is contextual and dependent on users, the explanation itself, and information users are interested in. The authors identify

properties of explainability by reviewing definitions of explainability from recent papers. The identified properties of explainability are used as objectives that evaluation metrics should achieve. The authors define explainability into components of interpretability, which is to be understandable to a human and fidelity, which is accurately describing model behavior. The interpretability of an explanation needs to be clear and unambiguous, able to be applied to many areas, and presented in a simple and concise manner. The fidelity of an explanation needs to be able to describe the whole process of the model and should be correct.

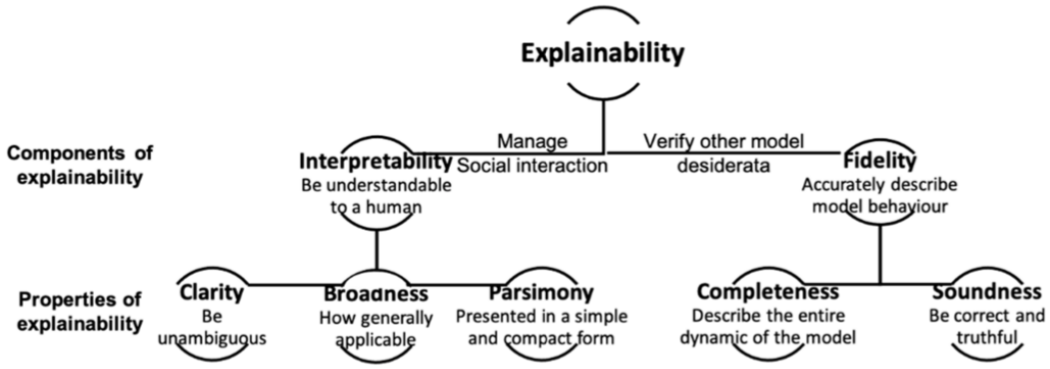


Figure 2. Definition of machine learning (ML) explainability and related properties (adapted from Reference [23]).

Explanations based on technology

Saliency methods highlights different parts in the data to understand classification tasks and this includes explanation methods like LIME & SHAP. Neural Network Visualizations are mainly used for neural network explanations and visualizes intermediate representations/layers of a neural network. Feature relevance methods studies the input features' relevance to and global effects on the target/output values. Machine learning models, such as deep learning models, are usually black boxes to users and difficult to understand. Knowledge distillation methods learn a simpler model, such as a linear model, to simulate a complex model and explain it.

Explanations based on context

Rationale explanation are about the "why" of an ML decision and provides reasons that led to a decision. Responsibility explanation concern about "who" is involved in the development, management, and implementation of an ML system. Data explanation focuses on what the data is and how it has been used in a particular decision. This type of explanation can help users understand the influence of data on decisions. Fairness explanation provides steps taken across the design and implementation of an ML system to ensure that the decisions it assists are generally unbiased, and whether an individual has been treated equitably. It can foster meaningful trust by explaining to an individual how bias and discrimination in decisions are avoided. Safety and performance explanation deals with steps taken across the design and implementation of an ML system to maximize the accuracy, reliability, security, and robustness of its decisions. Impact explanation concerns about the impact that the use of an ML system and its decisions has or may have on an individual and on a

wider society. This type of explanation gives individuals some power and control over their involvement in ML-assisted decisions.

Evaluating Explanations

The authors had two goals for evaluating explanations. The first is to compare available explanation methods and find preferred explanations from the comparison. The second is to assess if explainability is achieved in an application. The focus of the assessment lies in determining if the provided explainability achieves the defined objective. They argue that two factors determine understandability of an ML system: the features of the ML system and the human's capacity for understanding, so they based their evaluations on incorporating both factors.

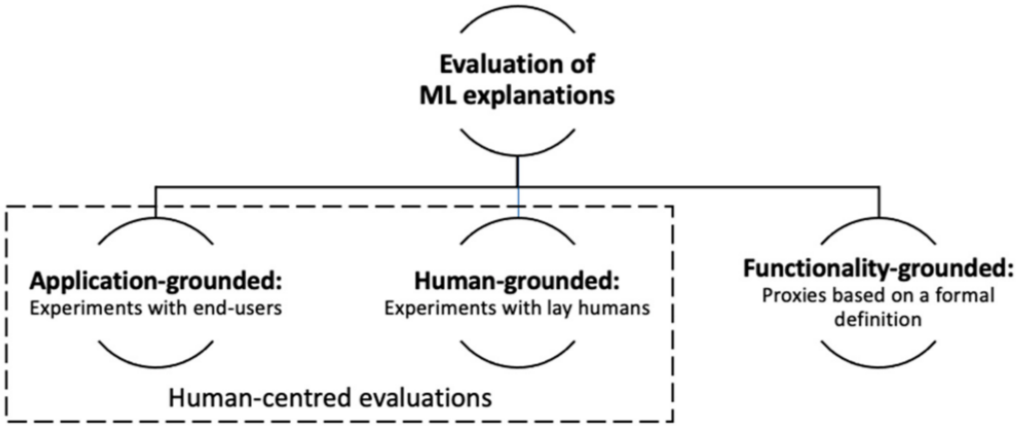


Figure 3. Taxonomy of evaluation of machine learning explanations.

The human's capacity of understanding takes the form of application-grounded and human-grounded evaluations. They define application-grounded evaluation as experiments with end-users. This kind of evaluation requires conducting end-user experiments using the explanation within a real-world application. They separate human-grounded evaluation as experiments with lay humans as opposed to end-users to assess the understandability of the explanations. It refers to conducting simpler human–subject experiments that maintain the essence of the target application.

The features of the ML system take the form of functionality-grounded evaluations. These are proxies based on a formal definition of interpretability to evaluate the explanation quality, e.g., the depth of a decision tree and does not require human experiments.

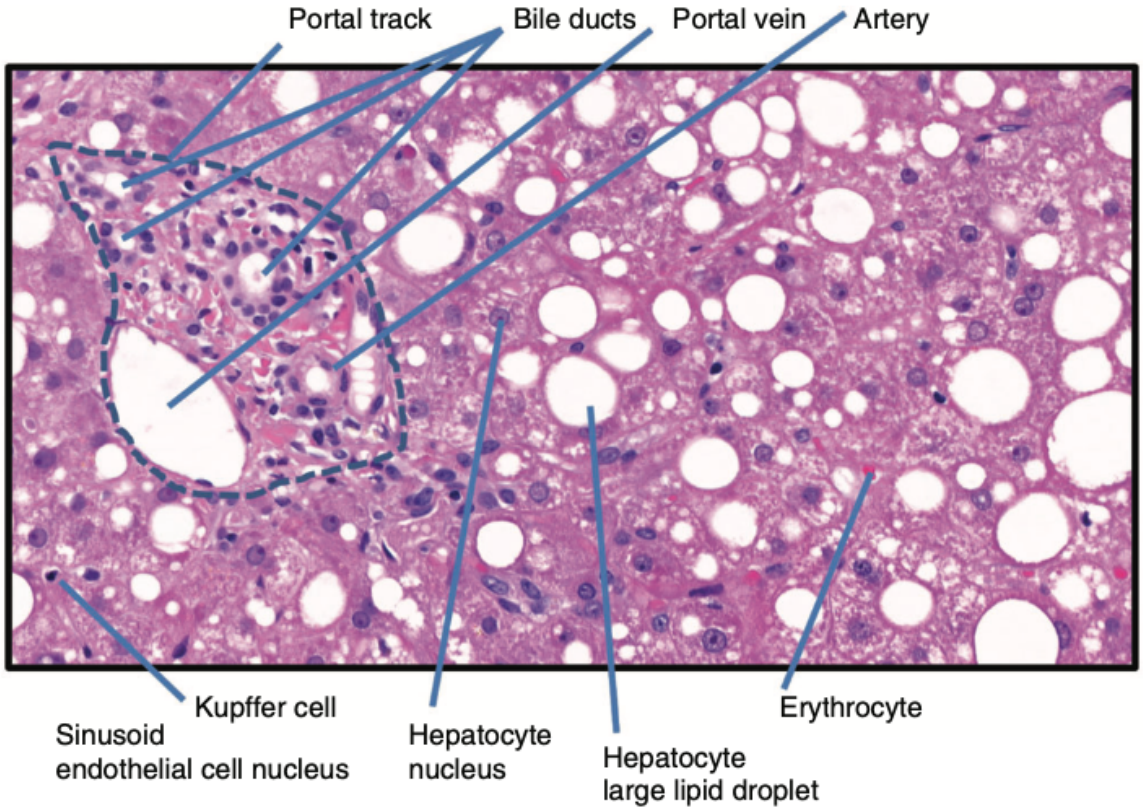
Finally, the authors boiled down human-centered evaluations into subjective and objective metrics. Subjective metrics can involve Subjective questionnaires, designed for users on tasks and explanations, and can be asked during or after task time to obtain user's subjective responses on tasks and explanations. Objective metrics involve objective information on tasks or humans before, after, or during the task time to assess behavior or task performance.

Concept of Causability

Another paper introduced the idea of causability which further reinforces the need for evaluating human understanding described in the previous paper. The authors refer to causability as the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use.

Complexity of Explanations of a Human Pathologist

To highlight how complex the casual understanding needs to be for explanations to become usable in the medical field, the authors asked for post-hoc and ante-hoc explanations from an experienced pathologist.



For the histology slide above, the authors asked the experienced pathologist to explain what he considered relevant in the histology slides, which he labeled the histology slide and explained certain facts such as liver biopsy with 10 evaluable portal fields, lobule architecture preserved, or liver cells arranged in regular plates one cell layer thick.

Then they asked the pathologist to explain the process and most relevant concepts in liver pathology, for an ante-hoc explanation. For liver pathology, we must describe specific features in the macroscopic evaluation of the histological section, the microscopic evaluation at low magnification, and the microscopic evaluation at higher magnification.

Future Avenues for Improving Causability

Supervised learning is very expensive in the medical domain as it is hard to get strong supervision information and fully ground-truth labels. Weakly supervised learning is umbrella for methods constructing predictive models by learning with weak supervision. So, the authors propose classifying whole slide images according to widely used scoring systems based on association with histomorphological characteristics and an overall predictive score using weak supervision.

Currently, models work on statistical free mode and use limited causal inference. We need to develop new visualization techniques that can be trained by medical experts, as they can explore the underlying explanatory factors of the data and formalize a structural causal model of human decision making and mapping features in deep learning approaches.

Finally, we need to incorporate three levels of causal hierarchy proposed by Judea Pearl to make sure causability measures ensure the "quality of explanations":

> Level 1: Association P(y| x) with the typical activity of "seeing" and questions including "How would seeing X change my belief in Y?", in our use-case above this was the question of "what does a feature in a histology slide tell the pathologist about a disease?"
>
> Level 2: Intervention P(y| do(x), z) with the typical activity of "doing" and questions including "What if I do X?", in our use-case above this was the question of "what if the medical professional recommends treatment X—will the patient be cured?"
>
> Level 3: Counterfactuals P(y_x| x , y) with the typical activity of "retrospection" and questions including "Was Y the cause for X?", in our use-case above this was the question of "was it the treatment that cured the patient?"

Evaluating XAI on a Multi-Modal Medical Imaging Task

Let's examine how evaluating explainable AI is done in an actual medical scenario. The authors of this paper chose to look at multimodal imaging, which refers to simultaneous production of signals for more than one imaging technique. For example, one could combine using optical, magnetic, and radioactive reporters to be detected by SPECT, MRI, and PET.

They make three key contributions. First, they conduct a systematic evaluation on a medical imaging task, that covers both quantitative and qualitative physician evaluation, and clinical requirements grounded computational evaluation on explanation faithfulness and plausibility. Then, they formulate and tackle the novel and clinically significant problem of multi-modal image explanation, which is the generalized form of single-modal image explanation. Finally, they propose the computational evaluation metric MSFI, which automates the human assessment process by incorporating the clinical patterns of modality prioritization and feature localization.

Faithfulness & Plausibility

The authors introduced the concepts of faithfulness and plausibility as the two main clinical requirements of explanation evaluation.

Faithfulness measures how accurately the explanation reflects the model's true decision process. It cannot be measured by human judgment or annotated ground truth encoding human prior knowledge, as humans have no idea about a model's internal decision process. Common evaluation methods include gradually erasing or adding features to input and measuring its effect on model performance. This concept is like functional-grounded explanations described in the first paper.

Plausibility is the users' assessment of how agreeable the explanation is with their prior knowledge of the task. It requires human annotated ground truth to reflect human prior knowledge on a given task, such as feature segmentation masks or bounding boxes. This is like application and human-grounded explanations described in the first paper and causability described in the previous paper.

Clinical Task

Using these two requirements of explanation evaluation, the authors focused on grading gliomas based on MRI, which could provide physicians indispensable information on patients' treatment plan and prognosis. They used the publicly available BraTS 2020 dataset (multi-institutional preoperative MRI scans) and trained a VGG-like 3D CNN with multi-modal 3D MR images to classify gliomas into lower-grade (LGG) or high-grade gliomas (HGG).

They conducted a physician user study to mimic explainable AI usage in clinical decision support, in which they asked neurosurgens to interpret, comment, and rate the generated 3D multi-modal heatmaps. In the figure below, we can see the MRI scans that were used, the heatmap showing areas of importance, and the question asked, "How closely the highlighted area of the heatmap match with your clinical judgement?"

Metrics used to Evaluate Explanations

The authors used the Modality importance (MI) metric which measures a model's overall importance of each modality as a whole using importance scores to indicate how critical is a modality to the overall prediction using Shapley values. This metric was looking at a global representation of importance, so they wanted to look at local importance, which they did using the Modality-specific feature importance (MSFI). This metric measures how well the heatmap can localize the modality-dependent important features in each modality. MSFI combines two ground-truth information: MI and modality-dependent important features. A higher MSFI score indicates a heatmap is better at highlighting important modalities and their localized features. If the feature localization annotations reflect human prior knowledge on the given task, then MSFI is a metric on explanation plausibility. Otherwise, if annotations reflect the intrinsic knowledge the model learned, MSFI can also be used as a metric for faithfulness.

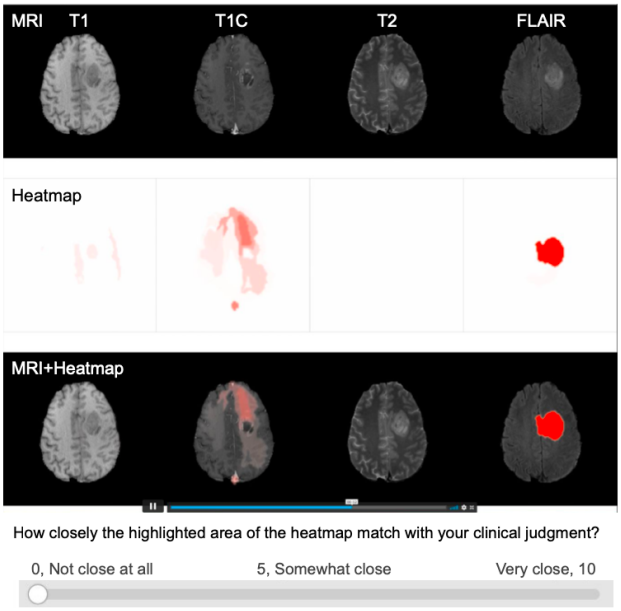


Figure 1: 3D heatmap (in video format) and questionnaire in the user study. Column: each MRI modality. Row: MRI, heatmap, and heatmap overlaid on MRI. Redness indicates the importance of that area for prediction.

Results

	MSFI (BraTS)	Stat. Sig.	MSFI (Synthetic)	MI Correlation	diffAUC	FP	IoU	Doctors' Rating	Speed (second)
Guided BackProp	0.48±0.33	NS	**0.49±0.21**	**0.80±0.27**	0.21±0.24	0.34±0.29	0.02±0.01	0.6±0.1	**1.7±1.1**
Guided GradCAM	0.50±0.36	⋆⋆	**0.42±0.29**	**0.81±0.26**	0.26 ± 0.25	0.37±0.31	0.02±0.02	0.1±0.0	2.2±1.4
InputXGradient	0.51±0.32	⋆	0.23±0.14	**0.87±0.16**	0.17 ± 0.12	0.40±0.30	0.08±0.05	0.1±0.0	**1.7±1.1**
DeepLift	0.54±0.34	⋆	0.22±0.23	0.53±0.45	0.19 ± 0.14	0.43±0.32	0.08±0.05	0.6±0.2	3.8±2.0
Integrated Gradients	0.48±0.31	⋆	0.22±0.19	0.73±0.39	0.17 ± 0.12	0.36±0.28	0.08±0.05	0.5±0.0	62±29
Occlusion	0.28±0.26	⋆⋆⋆	0.22±0.25	0.60±0.33	0.13 ± 0.15	0.18±0.19	0.03±0.02	0.6±0.2	989±835
Gradient Shap	0.48±0.31	⋆	0.22±0.19	0.53±0.40	0.17 ± 0.12	0.36±0.28	0.08±0.05	0.5±0.0	6.8±3.0
Feature Ablation	0.48±0.30	⋆⋆⋆	0.19±0.23	0.27±0.44	**0.30 ± 0.15**	0.35±0.28	0.05±0.06	0.4±0.4	74±23
Gradient	0.34±0.23	NS	0.19±0.13	0.47±0.16	0.05 ± 0.09	0.20±0.16	0.02±0.01	0.6±0.6	1.8±1.1
Shapley Value Sampling	0.38±0.24	⋆⋆⋆	0.10±0.10	0.47±0.65	**0.35 ± 0.04**	0.25±0.21	0.04±0.05	0.2±0.1	2018±654
Kernel Shap	0.28±0.25	⋆⋆	0.08±0.08	NaN	0.26 ± 0.16	0.18±0.20	0.06±0.08	0.1±0.0	194±100
Feature Permutation	0.23±0.26	NS	0.08±0.07	NaN	0.05 ± 0.05	0.13±0.18	0.05±0.07	0.1±0.0	14±2.2
Lime	0.24±0.21	⋆⋆	0.05±0.07	0.53±0.58	**0.37 ± 0.08**	0.14±0.16	0.05±0.06	0.1±0.0	341±181
Deconvolution	0.26±0.23	NS	0.04±0.02	0.73±0.39	0.11 ± 0.21	0.17±0.17	0.02±0.01	0.4±0.4	1.8±1.0
Smooth Grad	0.27±0.17	⋆	0.03±0.02	0.67±0.00	0.29 ± 0.25	0.16±0.12	0.02±0.01	0.7±0.1	12±6
GradCAM	0.04±0.03	⋆⋆⋆	0.02±0.02	NaN	0.16 ± 0.19	0.02±0.01	0.02±0.01	0.0±0.0	**0.6±0.3**

Table 1: Evaluation results. The table shows mean ± std for each XAI algorithm regarding different evaluation metrics on the test set. Metrics are in the range $[0, 1]$ (except for diffAUC and MI which is $[-1, 1]$), the higher, the better. Metrics for faithfulness and plausibility are marked with solid and dotted underline respectively, with bolded text indicating the top faithfulness performance for a metric. Stat. Sig. tests the correlation between MSFI (BraTS) score and the two groups of correct/incorrect predictions, with $\star$ indicates $p < 0.05$; $\star\star$ for $p < 0.01$; and $\star\star\star$ for $p < 0.001$; NS for not significant. "NaN" in MI is because the heatmap is not modality-specific and the correlation is not computable. Speed is the time spent to generate a heatmap.

Based on the table of results showing 16 post hoc heatmap algorithms and the values for each of the metrics gathered, the authors concluded that existing explainable AI algorithms failed to fulfill clinical requirements.

The algorithms were not faithful to the model decision process at the feature level (evidenced by the low and unstable MSFI on synthetic data and diffAUC), and users' plausibility judgment of the explanation was not indicative of the AI model decision quality (evidenced by the MSFI (BraTS) statistical test and its distribution visualization).

The poor and unstable explanation performance may lead to undesirable consequences in clinical settings. For example, in our user study, the authors observed doctors tend to assume the explanation is totally faithful to the model's decision process, therefore would take or reject the model's suggestion by judging the plausibility of the explanation. Given doctors' mental model on the assumption of totally faithful explanation, the evaluation for faithfulness should be put ahead of the evaluation for plausibility on model decision quality indication.

Explanation Trustworthiness

Finally, we will investigate explanation robustness, highlighted in the limits of explainability section, and if explanations can give reasons that are consistent across consistent input. The authors in this paper wanted to demonstrate potential problems with saliency maps using adversarial attacks. They used a public chest radiograph dataset, CheXpert, and a widely used baseline model, DenseNet to predict disease diagnosis. They used the following saliency map explanation algorithms: Vanilla Grad, Grad x Image, GradCAM, Guided-GradCAM, IG, SG and XRAI to show where the model looking at when making a prediction.

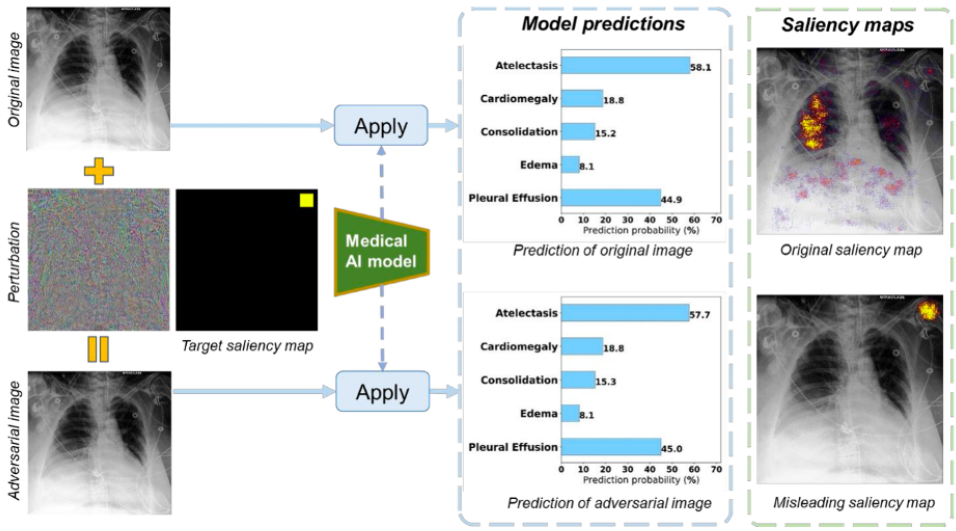


Figure 1. Our study illustrated by an example, where an adversarial image with human-imperceptible tampers the saliency map without changing the model predictions. The data and model used in this example are available at (https://github.com/DIAL-RPI/Trustworthiness-of-Medical-XAI).

The authors add some perturbations to the original image to create an adversarial image with a misleading saliency map. without affecting the model's predictions. By doing this process with all explanation algorithms, they show that the concern is universal.

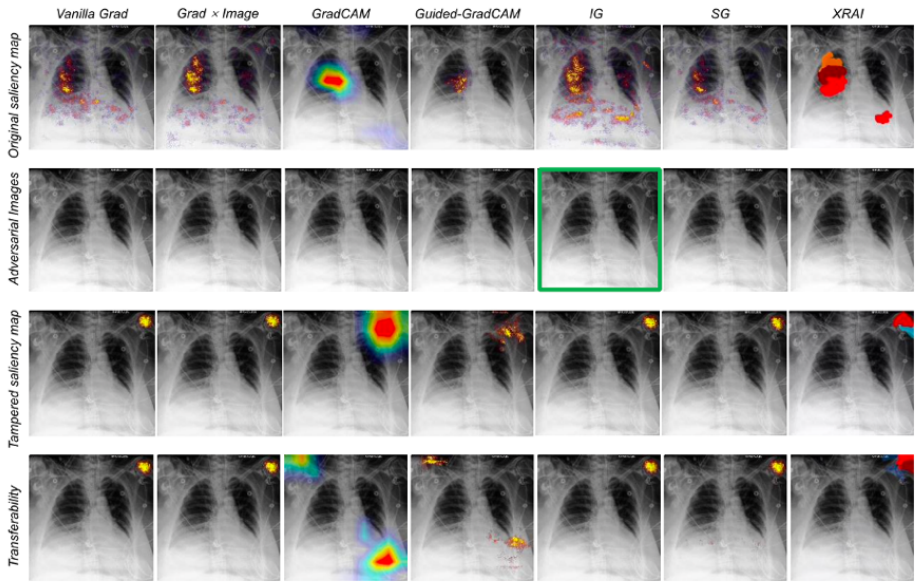
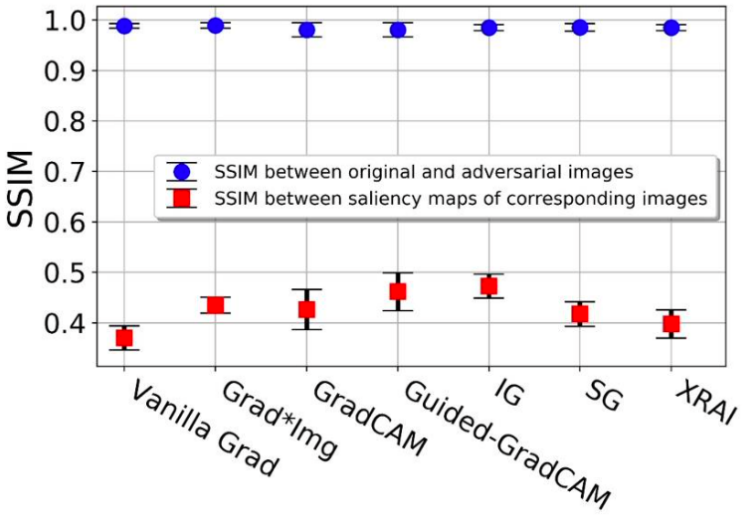


Figure 2. Collusion of explanation methods. ***First row****: saliency maps generated using different methods on the same original input image;* ***Second row****: generated adversarial images for each saliency visualization method.* ***Third row****: tampered saliency maps using the above adversarial images;* ***Last row****: saliency maps generated on the adversarial image highlighted by the green box in the second row.*

They elaborate further saying that ensuring model consistence isn't enough as they model prediction consistence on 200 chest radiographs from the CheXpert validation set. They show consistently, the adversarial images are near identical to the original images, but the saliency maps are very different between the original and misleading saliency maps.

The authors conclude that this is a problem because such perturbations could appear naturally in medical imaging due to numerous variations in commercial vendors, image reconstruction algorithms, patient characteristics, and so on.



Conclusion

From reading all these papers on the definitions of explainability, the need to expand it to include the human understanding, the evaluation of explainability in multi-modal medical imaging, and the limits of explainability, there are a few key takeaways.

The first major takeaway is that understandability of an ML system depends on two factors. First, it depends on the features of the ML system. We have seen this concept come up in various forms in multiple papers. This was referred to as functional-grounded explanations in the paper on defining explainability and to faithfulness of explanations in the paper on multi-modal medical imaging. Second, it depends on the human's capacity for understanding. This notion came up in all the papers that I have looked at, which highlights how important this is especially as a direction explainable AI needs to take. This concept was referred to as plausibility in the paper on multi-modal medical imaging, causability in its paper, and application/human-grounded explanations in the paper defining explainability.

The second major takeaway is that Solid quality-assurance measures are urgently needed to increase and validate the trustworthiness and robustness of AI explainability. There needs to be more regulations from bodies like the FDA to monitor the safety and effectiveness of systems in explainable AI.

Papers Selected

1. Beware explanations from AI in health care: https://www.science.org/doi/10.1126/science.abg1834

2. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics: https://www.mdpi.com/2079-9292/10/5/593

3. Causability and explainability of artificial intelligence in medicine: https://pubmed.ncbi.nlm.nih.gov/32089788/

4. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements?: https://arxiv.org/abs/2203.06487

5. Overlooked Trustworthiness of Explainability in Medical AI: https://www.medrxiv.org/content/10.1101/2021.12.23.21268289v1.full