Subjective evaluation of AI explainability methods and their applicability to chest x-rays

Author:

Samantha Escobar Martínez

Coordinator	Dr Karsten Tolle
Supervisor	Dennis Vetter MSc
Johann Wolfgang Goethe University	
Major	Computer science

Submitted: Monday 9 May, 2022

Contents

1 Introduction

The creation of more powerful machine learning models implicates the utilisation of more and more complex architectures and implementations, leading to a trade-off between transparency and performance (Lipton, 2018). This becomes all the more evident in deep neural networks, where the great abstraction caused by the high number of layers causes a sacrifice of interpretability/transparency for accuracy/performance (Selvaraju et al., 2017). While it may be tedious, it is still possible to manually solve the mathematical operations being performed under the hood by a perceptron or very simple neural network. Lipton (2018) delves into how to the simulatability of a model -the feasibility for a human to take the input data and manually go through every calculation to come up with a prediction within reasonable time- suggests transparency by empirically demonstrating a model is fully understood.

Machine learning models used for research nowadays utilise thousands of parameters and high dimensional data representations, making it unfathomable for humans to retain and recreate machine-made calculations manually. An unwanted scenario derived from this would be a black box that may deliver accurate results, while we remain ignorant as to how it's making the decisions we receive as output.

Interpretations of models help humans investigate the potential causality in correlations between data. Knowing how a model makes its decisions allows us to gain crucial understanding about its modus operandi, which in turn lets us debug our model more efficiently, corroborate made predictions and design improvements derived from it (Lipton, 2018).

Lipton (2018) emphasise the necessity for interpretability as a means to engender trust and a requirement for fairness, similar to how explicability is seen as an ethical principle necessary for the creation of turstworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) . It is

important to specify that trust is not to be seen as mere confidence in the predictive capabilities of a model but also as the willingness to relinquish control to the model. If we trust a model, we are more likely to agree relinquishing control to it.

An important caveat here is that the focus doesn't lie on how often a model is right but for which cases it is right. If the model tends to make mistakes in input space regions where humans also make mistakes and is accurate when humans are, it can be considered trustworthy. However, if the model makes mistakes for inputs humans can classify accurately, this shortcoming represent an added cost which may not warrant relinquishing control to the model, ergo a more present human supervision is needed (Lipton, 2018).

Selvaraju et al. (2017) refers to transparency and the interpretability to understand why a given prediction came to be as explicitly necessary for building trust on AI systems, highlighting its importance throughout the stages of evolution of artificial intelligence:

1. A system that is "weaker" than humans and considered unreliable, hence transparency is needed to identify a model's failures.

2. A system that is on par with humans and considered reliable (such as an image classification model with appropriate training) requires building trust to relinquish control and reduce direct human oversight.

3. A system is stronger than humans (for instance at chess or Go) and we want explainability in order to learn from the model.

Furthermore, the bigger the impact the decisions taken by a model ultimately have, the more trust is required, especially if the model's prediction does not concord with the opinion of a human expert of the respective field.

For example, if a model designed to identify tumours based on x-rays claims a patient has a

cancerous tumour requiring surgical intervention, whereas a radiologist disagrees and deems surgical procedures unnecessary one can greatly benefit from a better grasp on how the model came to such conclusion in order to refute or agree with its prediction more reliably. Similarly, we do not want to run into a paternalistic scenario where we unconditionally act as suggested by the model despite being oblivious as to how it operates and why it makes its decisions the way it does (Zicari et al., 2021); such a conundrum would also raise the question as to whom bears the responsibility when a fatal mistake is made either by a user (e.g clinician utilising predictive model to support assessment) or by the model (e.g considering the possibility of holding the model developers accountable).
The work presented in this report spans the implementation of a pre-trained deep neural network for image classification, the comparative evaluation of different explainability approaches as well as the integration and evaluation of a metric for quantitatively assessing the quality of provided explanations. Finally, the evaluation of explanation techniques is also performed with a real-world AI system.

2 Explainability methods

Throughout the implementation phase we utilised the deep neural network resnet 50, which is a variant of Resnet (He et al., 2016) with 48 convolutional layers along with one max pooling and one average pooling layer; such networks mitigate the problem of vanishing gradients and make use of residual connections to efficiently learn the identity. Our network was pre-trained on the ImageNet dataset, a visual databased frequently used in deep learning research consisting of over 14 million manually labelled images (Deng et al., 2009). For ease of reference, we shall focus most of the following examples on the same 224x224 pixel colour image (see figure 1) fed into our pre-trained instance of Resnet 50, which was correctly classified by the model as 'golden retriever'.

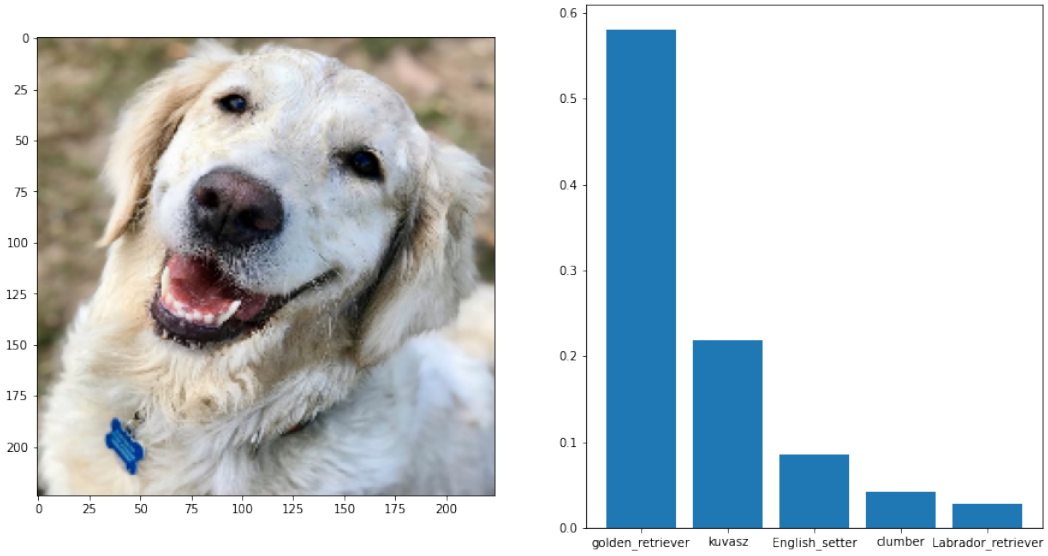


Figure 1: Resnet 50 model's top 5 prediction scores for image of class 'golden retriever'

The same image was explained utilising the following methods:

- LIME (Local Interpretable Model-agnostic Explanations)
- SHAP (SHapley Additive exPlanations)
- Saliency maps
- Integrated gradients (IG)
- Grad-CAM (Gradient-based Class Activation Mapping)
- XRAI (eXplanation with Ranked Area Integrals

2.1 LIME (Local Interpretable Model-agnostic Explanations

LIME is a model-agnostic approach introduced in Ribeiro et al. (2016), which proposes an explanation model learned locally to approximate a given prediction. The first step is to aggregate pixels into

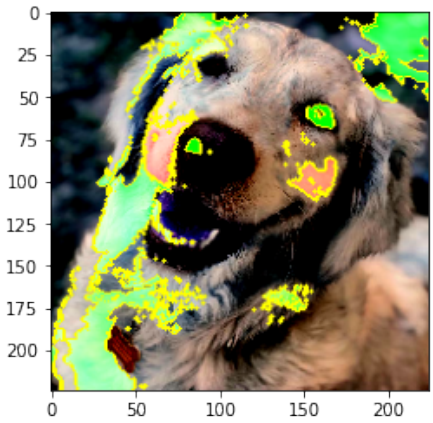


Figure 2: LIME explanation for predicted class 'golden retriever' showing the most influential superpixels for the prediction. Green superpixels increase the prediction contribution, while red superpixels reduce it (negative influence). [1]

superpixels to effectively delimit regions within the image in question. LIME trains a local surrogate model to approximate a given individual prediction through probing the model being investigated. In order to discern what regions of an image are the most influential for the underlying model's decision to predict it as belonging to a certain class, LIME generates a new dataset consisting of perturbed image samples (copies of the image with randomly deactivated superpixels) and the respective prediction outcomes when fed into the original model; an interpretable model is then trained to assess which regions of the original image are the most influential in the model's predictive decision. In figure 2 we can observe the most influential superpixels highlighted in an image predicted as 'golden retriever', where the green superpixels have the largest contribution increase to the prediction, while the red ones represent he largest decrease.

[1]LIME explanation created with the LIME python package https://pypi.org/project/lime/

2.2 SHAP (SHapley Additive exPlanations)

This method leverages the notion of Shapley values from coalitional game theory -where one quantifies the contribution of every player to a total payout- to explain the prediction of an instance by computing the contribution of each feature to the prediction. As with LIME, the image in question is not represented on the pixel level but as an aggregation of superpixels.

A positive Shapley value indicates a superpixel contributes to increasing the explained prediction, while a negative value decreases it. A series of coalitions is randomly sampled; a coalition being a possible configuration of the image in question with certain superpixels turned on while others are turned off (removed from the image / replaced with black pixels). After getting the predictions for each one, the weight of each coalition is computed, giving larger weights to the coalitions with just one or most features turned on or off, as they tell us about the impact of individual features in isolation. Conversely, a coalition where around half of the features (superpixels) are present, we won't gain much insight on an individual feature's contribution, thus it is assigned a very small weight. The fundamental difference in contrast to LIME lies here: whereas SHAP assigns weights to the sampled instances according to the weight the coalition wold get in the Shapley value estimation (small and large coalitions are assigned the largest weights as they tell us the most about the contribution of individual features through their isolated presence / absence), the weight value assigned by LIME is directly proportional to the number of present features in a given coalition e.g. a sample where only one superpixel is present would have a very large weight when using SHAP and a very small weight when using LIME.

Figure 3 shows us which regions of the image SHAP computed a big contribution impact (both positive and negative) on its prediction. Having not only the top class but also the following 3 labels and their explanations displayed next to each other helps elucidate not just why an image is labeled

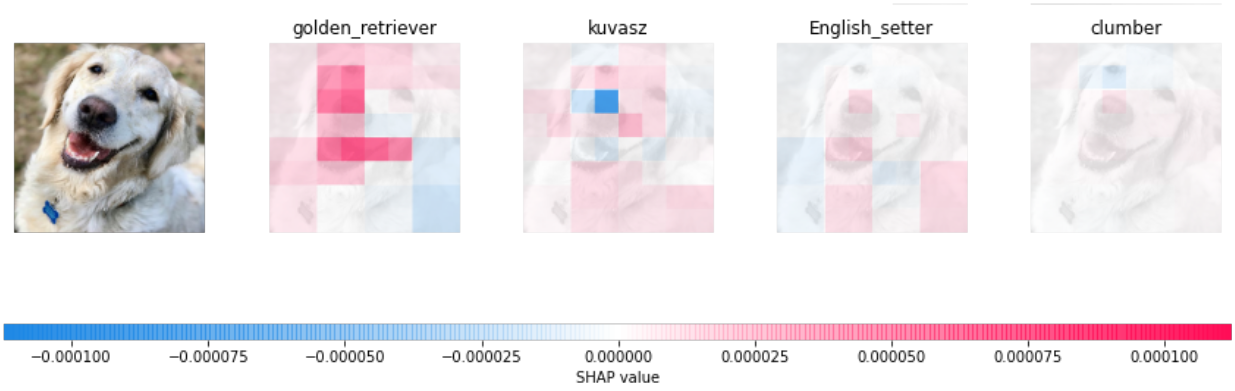


Figure 3: SHAP explanation for the top 4 predicted classes. Blue superpixels detract from the prediction, whilst red superpixels increase the prediction contribution.[2]

as a certain class but also why it's not labeled as a different one.

2.3 Saliency maps

Saliency maps are a widely used method to explain the final classification decision of a convolutional neural network. This form of pixel attribution quantifies the contribution of every pixel in the image to its ultimate prediction outcome by establishing a directly proportional relation between the pixel intensity (its produced output is visualised as a heatmap) and said pixel's relevance towards ultimately labelling the input image as belonging to the predicted class.

Simonyan et al. (2013) introduces an approach to quantify these pixel attributions in a way akin to the common computation of a gradient used in backpropagation: After performing a forward pass on a given image, we compute the gradient of the class score of interest with respect to the input pixels and visualise the gradients as observed in figure 4.

[2]SHAP explanation created with the SHAP python package https://pypi.org/project/shap/

[3]Saliency map explanation created using Tensorflow keras and Uzman Riswan's implementation https://usmanr149.github.io/urmlblog/about.html

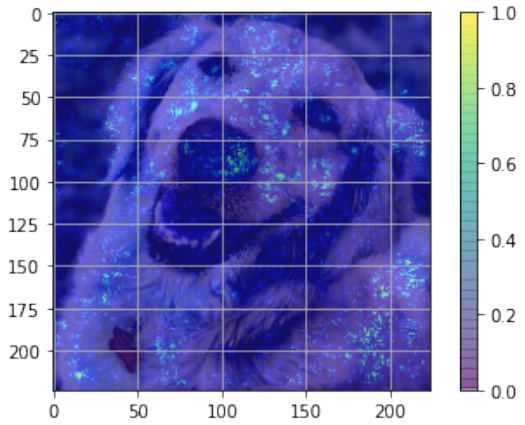


Figure 4: Saliency map for predicted class 'golden retriever'. Higher intensity gradient visualisations in the heatmap represent a bigger contribution to the prediction. [3]

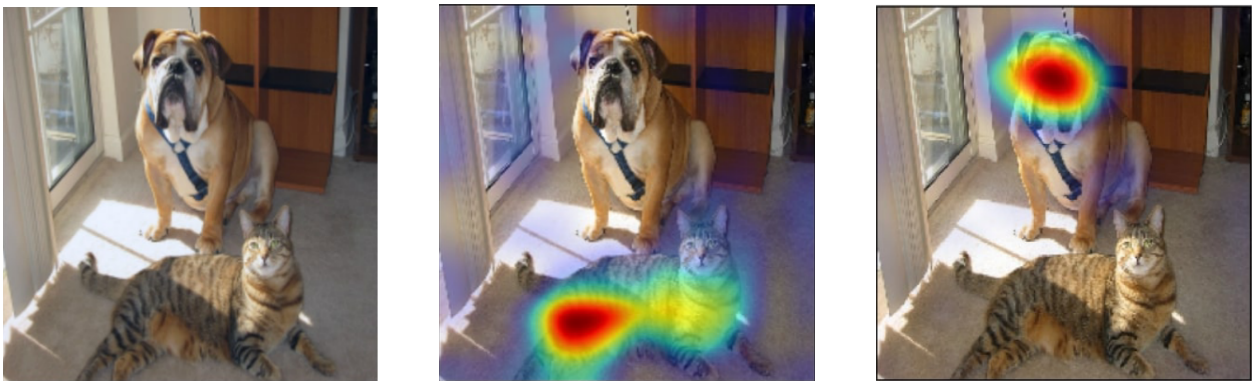
2.4 Grad-CAM (Gradient-based Class Activation Mapping

Selvaraju et al. (2017) sets specific criteria for what they deem a good visual explanation:

- The explanation must be class-discriminative, meaning the category can be localised within an image. For example, in figure 5 we can see how we apply pixel-space gradient visualisation in the area of the image where the cat is being highlighted when explaining 'cat' (the same for explaining 'dog'), making it easier to fathom why such a prediction occurred
- The explanation must be high-resolution, meaning it is capable of capturing fine details.

Based on these criteria, they developed Grad-CAM. The goal of this method is to produce visual explanations for decisions from CNN-based models for tasks such as structured outputs, visual question answering and reinforcement learning; without the necessity of re-training or architectural changes.

A localisation map is produced to highlight important regions in the image for the resulting prediction. A major aspect that differentiates this method from saliency maps is that a high gradient doesn't



(a) original image (b) Grad-CAM 'Cat' (c) Grad-CAM 'Dog'

Figure 5: Grad-CAM class-discriminative visualisation of classes 'Cat' and 'Dog' (Selvaraju et al., 2017).

necessarily indicate a feature being very important. Whereas with saliency maps one only aims to visualise the gradients, without any distinction of relevance between them, Grad-CAM looks not only into the gradient but also the activation of the last layer.

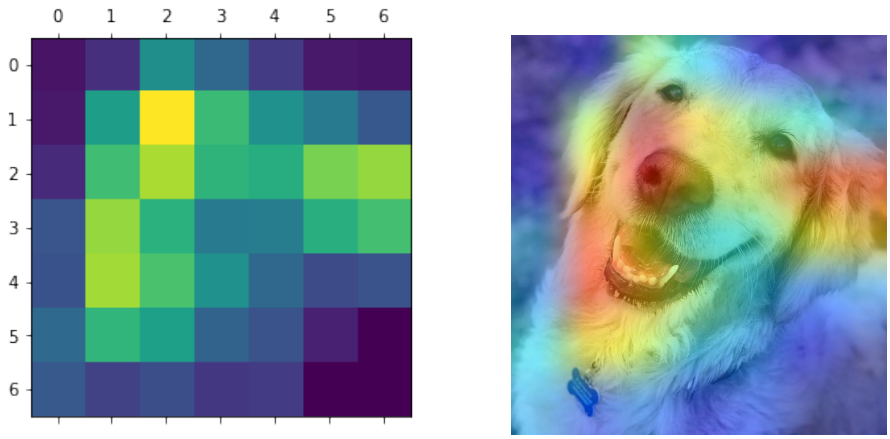
Figure 6 shows us the resulting visualisation of applying Grad-CAM on our example image by scaling. Having the original heatmap helps clarify (in ensemble with the final visualisation) which areas are of major importance to the classification, such as eyes, ears and mouth.

2.5 Integrated gradients (IG)

Parting from the initial motivation to facilitate the empirical evaluation of attribution techniques, integrated gradients is a feature attribution method that aims to help the user understand on which features a neural network relies for making its predictive decisions and quantify feature importance (Sundararajan et al., 2017).

Integrated gradients revolve around missing features. One begins with a baseline image (typically a completely black image or random noise) of the same dimension as the sample in question, which is

[4]Grad-CAM explanation using Tensorflow keras https://keras.io/examples/vision/grad_cam/



(a) Grad-CAM heatmap for the prediction 'golden retriever'

(b) Scaled heatmap overlaid on input image

Figure 6: Scaling and juxtaposition of heatmap for explanation visualisation of class 'golden retriever'. [4]

Figure 7: Image samples obtained through linear interpolation between baseline and image to be explained (class 'golden retriever')

then linearly interpolated with the image to be explained. Figure 7 shows how we create a series of images starting from a baseline until we reach complete opacity of the image to be explained (on the right). We want to investigate what is the minimum number of features (in this case it's represented by the interpolation sample with the lowest opacity of our target image) we require for the gradients to saturate, reaching a point in the linear interpolation process from which the prediction score will not increase regardless of the opacity increase of the target image.

[5]Integrated gradients explanation Using Tensorflow keras https://www.tensorflow.org/tutorials/interpretability/integrated_gradients

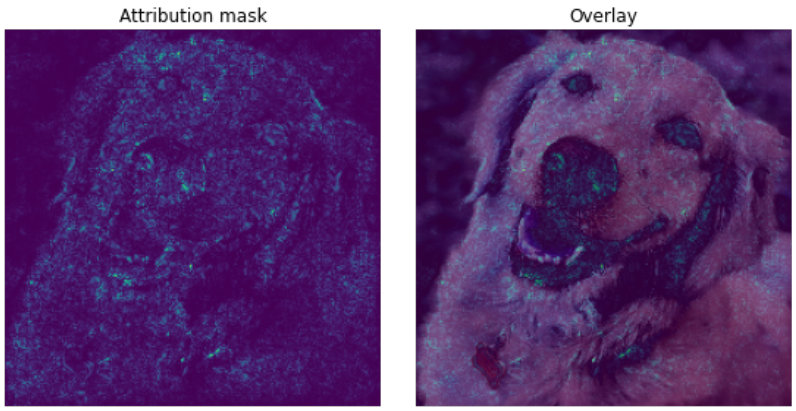


Figure 8: Attribution mask of the integrated gradients (isolated and overlaid on target image): higher intensity values represent a larger importance/contribution to the prediction. [5]

2.6 XRAI (eXplanation with Ranked Area Integrals)

A problem with saliency maps and other individual pixel attribution methods is that the produced explanations are difficult for users to interpret, particularly in cases where the resulting heatmap is sparsely spread out across the entire image. For instance, figure 4 suffers from this setback, as the resulting heatmap makes it difficult for humans to clearly visualise what parts of the image led to the ultimate prediction.

XRAI aims to solve this by performing the attributions on regions of the image instead of individual pixels. The image is separated into regions based on neighbouring pixel dissimilarity; then the pixel-level attribution within said regions are aggregated to visualise areas that positively or negatively impact a given prediction.

The algorithm can be superficially described as shown in figure 9:

1. **Pixel-level attribution:** Integrated gradients with a black and a white baseline is are used to perform pixel-level attributions for the input image

2. **Oversegmentation:** Separately from model attribution, the image is segmented into regions

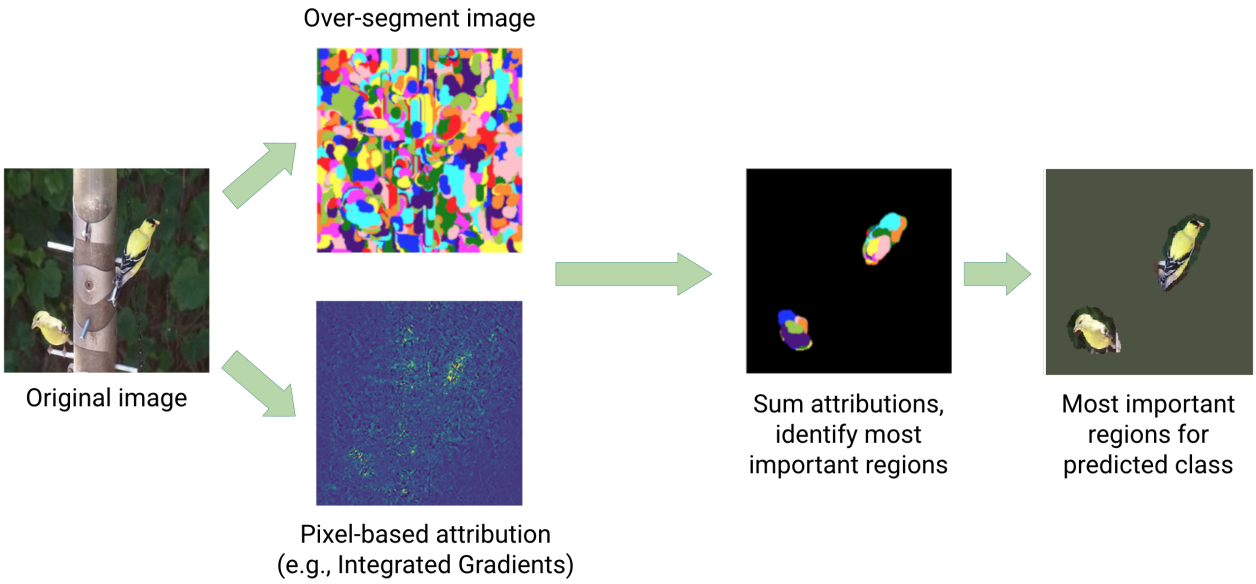


Figure 9: Identification of most influential regions for a prediction (Kapishnikov et al., 2019)

(Felzenwalb's graph-based representation (Felzenszwalb and Huttenlocher, 2004))

3. **Region selection:** For each region, the individual pixel attributions within it are added up. The segments are ultimately rank-ordered from most to least positive based on the attribution sums. Additionally, the method allows to create an image composed of only the top n% most influential regions to the prediction, as can be seen in figure 10.

3 Qualitative assessment of explanations

Our intention is to examine the efficacy of the mentioned explainability techniques. Since there is no globally adopted metric for the numerical evaluation of explanation methods, we will use the scale created in Holzinger et al. (2020): Holzinger et al. (2020) proposes the system causability

[6]XRAI explanation created with the saliency python package https://pypi.org/project/saliency/

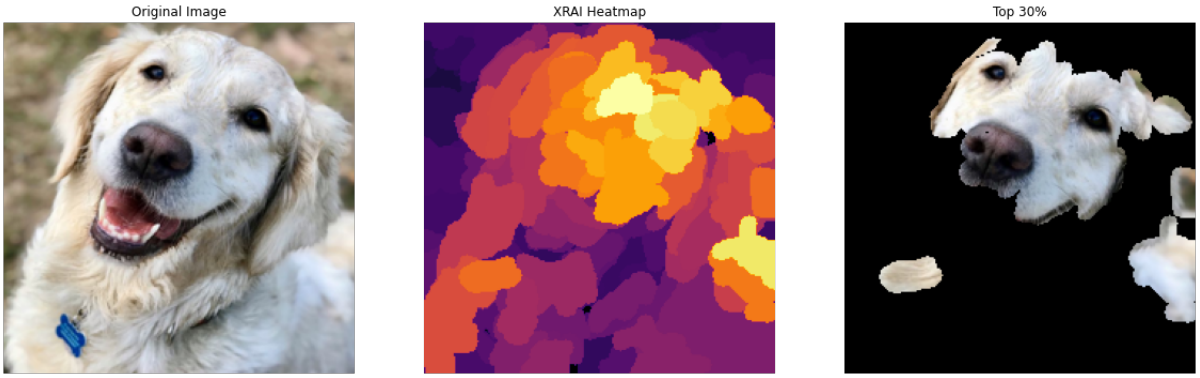


Figure 10: XRAI output on example image, including the top 30% most influential regions to the prediction 'golden retriever'.[6]

scale (SCS) to evaluate the quality of explanations to facilitate the comparison and analysis of explainability methods. Holzinger et al. (2020) emphasises establishing a clear dichotomy between explainability and causability when investigating a machine statement (prediction): Explainability is referred to as the ability to highlight decision-relevant parts of machine representations and models; i.e. in our case this would refer to the parts of the image which contributed the most to the given prediction.

Causability is referred to as the extent to which the provided explanation to a machine statement leads to a human user gaining effective and satisfactory causal understanding about the model's decision.

3.1 Data collection

The introduced system causability scale (SCS) (Holzinger et al., 2020) aims to quantitatively investigate and determine whether and to what extent an explanation is suitable and efficacious for its intended purpose. Said scale consists of the following ten statement questionnaire in which every

statement is to be answered with one out of five potential answers ranging from *strongly agree* to *strongly disagree*:

1. I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. I understood the explanations within the context of my work.
3. I could change the level of detail on demand.
4. I did not need support to understand the explanations.
5. I found the explanations helped me to understand causality.
6. I was able to use the explanations with my knowledge base.
7. I did not find inconsistencies between explanations.
8. I think that most people would learn to understand the explanations very quickly.
9. I did not need more references in the explanations: e.g., medical guidelines, regulations.
10. I received the explanations in a timely and efficient manner.

Each answer is mapped to a score from 1 to 5 (1 being *strongly disagree* and 5 *strongly agree* and the sum of the scores is divided by the maximum possible score (50).
Aiming to obtain a more thorough insight on the methods' performance, we opted to have a group of people assess the models with the system causability scale through a survey.

Question	Saliency map	LIME	SHAP	Grad-CAM	IG	XRAI
01. Factors in data	2.5	2.375	3.125	2.875	3.125	**3.625**
02. Understood	2.75	3.375	2.875	2.75	3.125	**3.625**
03. Change detail level	**2**	1.875	**2**	**2**	1.875	**2**
04. Need teacher/support	**3.375**	3.125	2.875	3.125	3.125	**3.375**
05. Understanding causality	3	3.375	3.625	3.375	3.375	**3.75**
06. Use with knowledge	2.75	**3.25**	3.125	3	2.875	**3.25**
07. No inconsistencies	2.625	2.25	**3.25**	2.75	2.75	3
08. Learn to understand	2.375	**3.5**	3	3	3.125	3.25
09. Needs references	2.625	2.875	2.875	**3**	2.875	**3**
10. Efficient	2.125	2	2.125	2.125	**2.25**	**2.25**
SCS = $\frac{\sum ratings}{50}$	0.522	0.6	0.577	0.56	0.57	**0.6225**

Table 1: System causability scale scores based on the answers on a group of 8 people from different areas in the STEM field.

3.2 Results

Although the idea behind following a standardised metric to assess the quality of explainability methods sounds promising in theory, applying the system causability scale in the context of our image classification explanations proves less fruitful and more complex.

Tables 1 and 2 show us how even the best scoring method (XRAI) performed very poorly when using this metric. One possible shortcoming may lie on the questionnaire itself and not only on the explanation methods, as most of the participants who answered the questionnaire stated that the questions were unclear and very ill-suited for the explanation methods at hand.

3.3 Discussion

One of the main contributions of this work is to empirically test the efficacy of explainability methods: While most publications that introduce such methods focus on their technical capabilities

Question	Saliency map	LIME	SHAP	Grad-CAM	IG	XRAI
01. Factors in data	2.428	2.375	3.125	2.875	3.125	**3.625**
02. Understood	2.625	3.375	2.75	2.5	3	**3.625**
03. Change detail level	**1.5**	1.375	**1.5**	**1.5**	1.375	**1.5**
04. Need teacher/support	**3.375**	3.125	2.875	3.125	3.125	**3.375**
05. Understanding causality	3	3.375	3.625	3.375	3.375	**3.75**
06. Use with knowledge	2.625	**3.125**	3	2.875	2.75	**3.125**
07. No inconsistencies	2.375	2	**3.125**	2.5	2.625	2.875
08. Learn to understand	2.375	**3.5**	3	3	3.125	3.25
09. Needs references	2.5	2.75	2.75	**2.875**	2.75	**2.875**
10. Efficient	1.5	1.375	1.5	1.5	1.625	**1.75**
SCS = $\frac{\sum ratings}{50}$	0.487	0.527	0.545	0.522	0.5375	**0.595**

Table 2: A modified version of the system causability scale where the participants were able to mark each questionnaire statement as "unsuitable/inapplicable" which was valued with a score of 0.

Method	Execution time (in seconds)
Saliency maps	2.251
LIME	29.886
SHAP	42.105
Grad-CAM	**0.2111**
IG	3.288
XRAI	8.479

Table 3: Time taken to execute each explanation method on the target image

and improvements over other existing methodologies, their efficacy and applicability remain mostly hypothetical. With this survey we intended to corroborate and investigate how much clarity such methods truly deliver and how intuitive they are.

- **Saliency maps:** An unintended advantage of this method was that it helped several participants familiarise themselves with the concept of explainability methods in general (transitively facilitating the understanding of the remaining explanation methods) by reportedly providing a theoretically simple and clear association between pixel intensity in the resulting heatmap and prediction contribution. Additionally, having a low execution time can be beneficial for future work where the technique is applied on a large pool of images for further investigation. That being said, the caveat must be made that this approach serves more as a way to visualise and highlight which pixels are relevant for the model's prediction without elaborating further into why this is the case.

- **LIME:** Highlighting the superpixels with the largest positive prediction contribution provides a clearer and more intuitive causal understanding of the prediction in contrast to individual pixel attributions, though the regions of the largest negative contributions are not as insightful to users given that once again, it is only a localisation of regions of influence without any further information explaining why this is the case.

  An additional drawback / source of uncertainty when using this method lies in its random element. As exemplified in figure 13, the output of the method can vary due to the stochastic nature of the process of superpixel perturbation to the point where discrepancies may occur such that areas that were marked as having a large positive contribution in one explanation impact appear as part of a superpixel with a large negative contribution in another explanation instance. Another drawback is the high execution time required, which may be a result of the necessity of creating a surrogate model to explain a single prediction.

- **SHAP:** Participants report that the contrastive way in which the explanations for the top four classes for a single image (see figure 3 are presented is of considerable benefit for causal understanding, tackling the exact issue experienced when looking at LIME explanations. A better grasp of the most confident prediction can be gained by observing what areas of the image detract from its confidence and more importantly, by contrastively looking at what areas were relevant for the other top classes (which helps fathom why these are not the most confident prediction). It was pointed out by participants that the way in which colour intensity is used to visualise the magnitude of the prediction contribution makes the interpretation of the explanation less clear as opposed to the LIME approach, where it is only discerned between the most influential positive and negative superpixels without further detail of their contribution magnitude. As with LIME, the high execution time required for this method -which is likely caused by the generation of all the possible coalitions- makes this method ill-suited for working with a larger pool of images.

- **Grad-CAM:** A strength of this method was that according to some participants, it is more intuitive and easier to comprehend when shown regions that contribute to a prediction as opposed to individual pixel attributions since this way one can easily make an association between the highlighted area and a relevant component of the target image (snout, mouth, ear). However (and particularly evident without the help of the original, unscaled heatmap), the visualisation of the contributions overlaid on the target image appears convoluted, making it difficult to identify and tell the most significant areas apart. A possible cause for this difficulty is that the target object (dog) takes up the vast majority of the area of the image, which may lead to most of the image being categorised as yielding relevant contributions and indirectly making it harder to visually discern the attribution areas. As this method boasts the fastest execution time, future work could entail a more in-depth assessment of this method utilising a

larger pool of images, such as with images where the target object does not take up so much space of the image.

- **Integrated gradients:** One could make the claim that the attribution mask of figure 8 provides a very clear visualisation of what pixels are important for the model's prediction due to the amount of detail (almost a silhouette of the target object) conveyed by it. Conversely, one could also claim that it is this amount of detail that makes for a poor explanation, since by depicting so much of the image as having a large contribution impact, one cannot discern particular areas that explain why it was classified as. If an explanation gives off the impression that everything (or an overwhelming part of the image) is important for its classification, it all inadvertently loses relevance by extension.

- **XRAI:** Several participants found this to be the most intuitive and effective method due to highlighting important regions as well as providing an isolated sample of the most significant areas, which are easily conceptually associated with target object components(eyes, snout). Nevertheless, it could be argued that the presented threshold of showing only the top 30% in terms of the contributing segments causes this method to also suffer from the same loss of relevance we encountered when applying integrated gradients; considering too much information as relevant causes the explanation as a whole to lose clarity. This value can easily be tweaked when generating explanations in future work, in turn enabling a user to further narrow down which segments are the most significant for the model's decision, which would grant a vast causal understanding of the given prediction. While this method's execution time may prove too long for being applied on a larger pool of data, it is significantly faster than the other methods that create visualisations through areas of attribution (SHAP and LIME), which were the approaches with the most clarity according to the participants. Hence, it would be

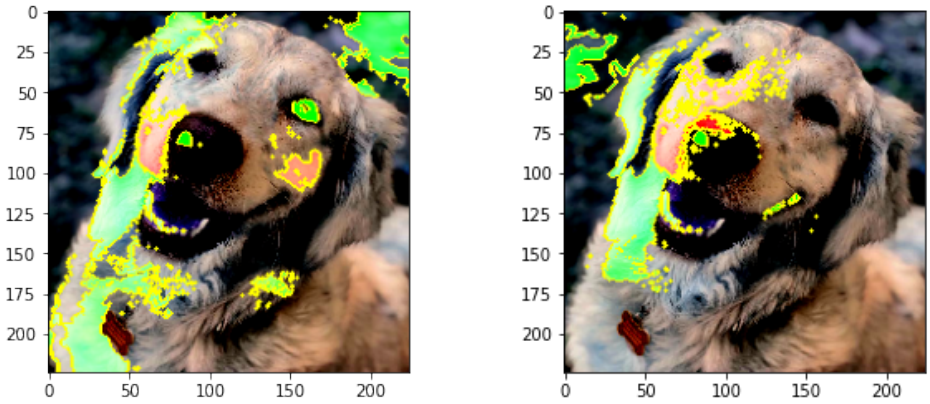


Figure 11: Two different LIME explanations of the same input image

interesting to experiment further with this method, for example by generating visualisations of the top 10%, 20% and 30% most influential regions for the model's prediction.

3.4 Limitations

The first issue arises with the fact that Holzinger et al. (2020) does not delve any deeper into what is meant with every statement of the questionnaire, leading to uncertainty and room for misinterpretation (or multiple interpretations of the same statement). While we intentionally opted to take a textual interpretation of every statement without modifying its message to fit the context of our experiments with the goal of maintaining objectivity, it quickly became evident statements such as number 3 "I could change the level of detail on demand" was ill-suited in the context of our work. It must also be taken into account that the conducted scoring of the explainability methods using this scale cannot be deemed sufficiently comprehensive, as each method was evaluated based solely on one example, thus hinting at the value of performing a more thorough analysis of the quality of every method (such as with a larger pool of samples, users from different fields of expertise participating, etc.) in future work.

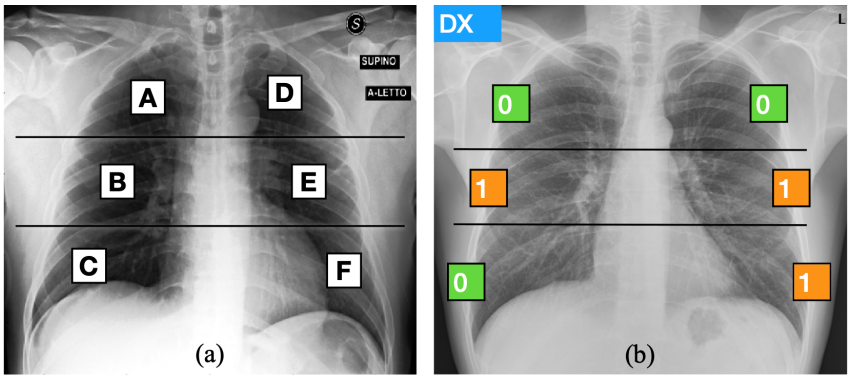


Figure 12: (a) Example chest x-ray image separated into 3 regions per lung. (b) A different example chest x-ray image with damage scores assigned to each pulmonary region (Signoroni et al., 2021)

4 Use case

Having gained a better understanding of explainability methods we wanted to investigate how well they apply and translate to a concrete, actual use-case. For this we look into a predictive model with its own explanation method to see if the aforementioned techniques could be applied on it and further examine how this explainability technique works and why it had to be created. Signoroni et al. (2021) introduces a model to quantify regional pulmonary compromise on COVID-19 patients using chest x-ray images as input.

The model operates by first isolating and segmenting the lung area of the chest x-ray image in question, which is then aligned and centered. This is followed by dividing each lung into three regions and assigns a score (whole number) between 0 and 3 to represent the level of compromise in said area, as can be seen in figure 12.

A novel post hoc method was created to visualise and explain the model's predictions.

We utilised the example of figure 13 with the system causability scale and obtained the results observed in table 4

From a subjective evaluation and by comparing it to the other approaches, we observed that this

Question	BrixIA
01. Factors in data	5
02. Understood	2
03. Change detail level	1
04. Need teacher/support	2
05. Understanding causality	2
06. Use with knowledge	2
07. No inconsistencies	1
08. Learn to understand	1
09. Needs references	1
10. Efficient	1
SCS = $\frac{\sum ratings}{50}$	**0.36**

Table 4: System causality scale scores for the BrixIA explanation method.

method appears to suffer from some of the same difficulties faced with the other explanation methods, which diminish its quality and effectiveness. In particular, the method could more aptly be called a description method instead of an explanation, which may possibly cause the user -a radiologist- to be put in a position where the model takes a paternalistic role, stating that the given colour-coded zones represent the magnitude of pulmonary compromise, yet it is never elaborated as to why this is the case.

It is also important to mention that this explanation method appears to be under development, as it was not included in the official publication of Signoroni et al. (2021). While demo code was provided upon request by the authors, there's no documentation nor further details available on how to implement or interpret the output of the explanation methods. For instance, there's no official explanation as to what the colour intensity of an area in the explanation image may represent, though based on the functioning of other methods one could speculate that the colour intensity is directly proportional to the prediction contribution.

An interesting claim made in Signoroni et al. (2021) is that their model cannot be explained using common post hoc visualisation methods like the ones we presented due to technical incompatibility.

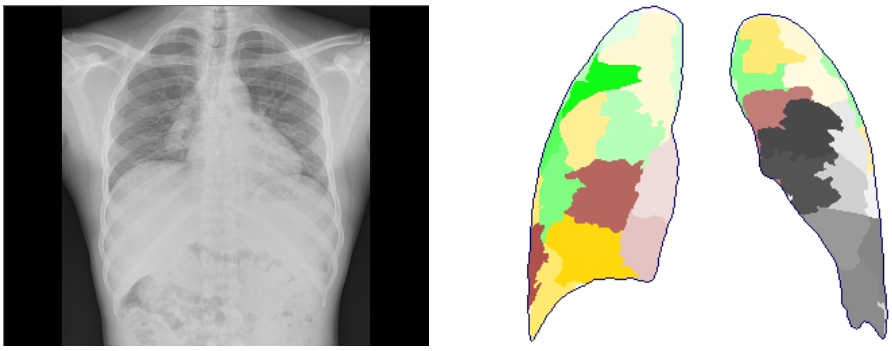


Figure 13: Original chest X-ray image and resulting explanation visualisation

As there was no further detail mentioned regarding this limitation, we decided to verify this claim by attempting to apply the LIME and SHAP methods on it.

A great obstacle that hindered the implementation process is the discrepancy between the state of the code mentioned in Signoroni et al. (2021) and what is actually available on their open-source repository. While the paper mentions in repeated occasions how their code is fully functional and accessible on their repository as well as putting an emphasis on its ease of use, these claims appear to apply mostly to the predictive model for generating scores of pulmonary compromise; the code for generating explanations for such predictions isn't readily available and was only obtained from the developers upon request and under the notice that it isn't complete and requires considerable additional configuration work from our end to properly function. It is also important to mention that in our experiments, generating a single explanation using this model led to reoccurring crashes, as the available 12.69GB RAM of the instance of Google Colab running did not suffice to create the explanation as proposed in the code sent by the developers.

The main difficulty encountered when attempting this was the fact that the BrixIA model has a fundamental difference in operation in contrast to the classification setting we used for visualisation above, as it produces six predictions of lung damage magnitude (one for each pulmonary region) per input chest x-ray. Additionally, the way in which the model separates each lung into three regions

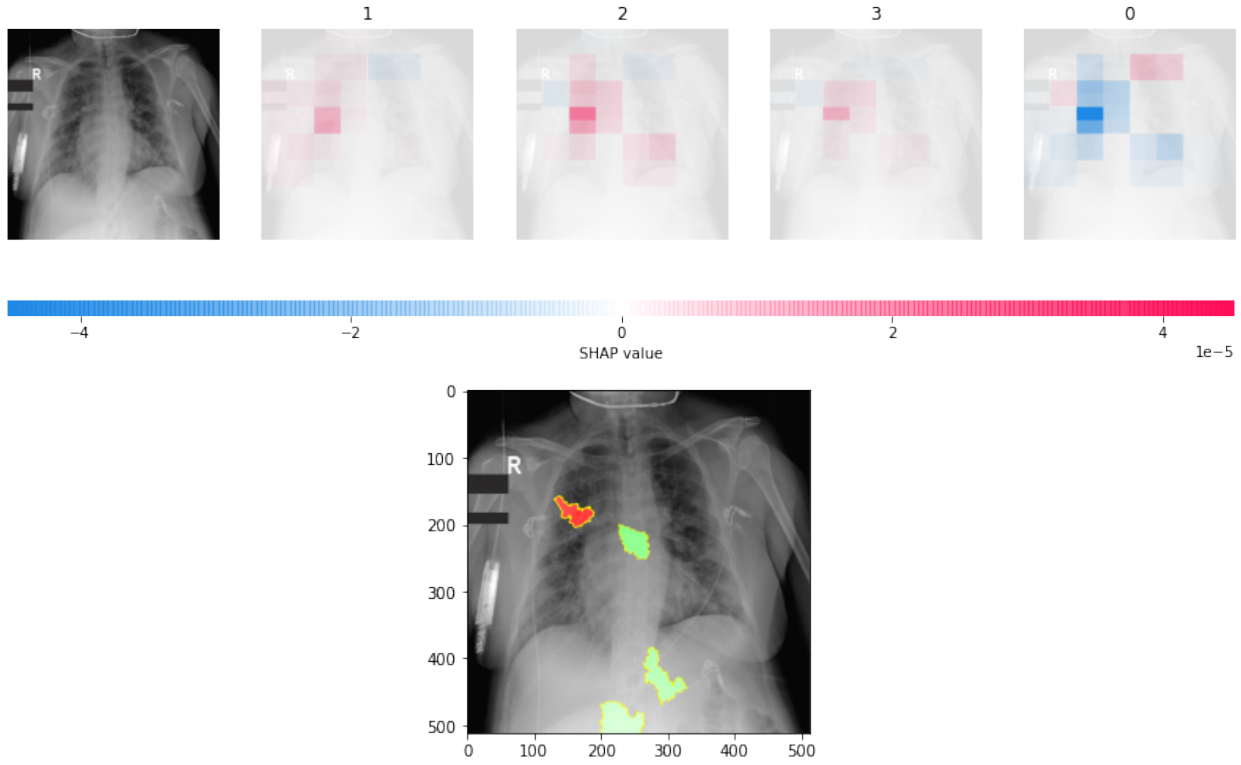


Figure 14: SHAP and LIME explanations for chest x-ray image where the damage prediction for the top left pulmonary region is 1.

and the spatial representation of such is inaccessible to the user, meaning there is no control neither further clarity on how they are designated.

Due to these difficulties, our attempt at applying the aforementioned explanation methods consisted in taking a chest x-ray image as input and using the BrixIA model to obtain a damage prediction a priori and generating explanations based only on the damage magnitude prediction of one pulmonary area.

Figure 14 helps exemplify how the implementation specifics of the model hinder the application of commonly used explainability methods. While we can see the highlighting of certain areas in the image to indicate positive prediction contribution, we cannot firmly establish causality to this

correlation, for we are performing an explanation of a single region with the entire image, making our approach unreliable.

We can see these inconsistencies manifested in the fact that areas other than the top left pulmonary region are highlighted to indicate positive or negative prediction contributions. In theory, pixels outside the region ought not to have an influence on the model's classification decision, which may suggest that the model is not functioning as originally envisioned or that the presented explanations are erroneous. Additionally, the input of a knowledgeable expert in the field of radiology would be beneficial (if not outright necessary) to objectively assess the explanations.

5 Conclusion

In retrospect, the main contribution of this work can be seen as an empirical demonstration and further corroboration on not just the importance of explainability but also on the facts that there is no global consensus on what the terms explainability and interpretability mean as well as how to assess it. Both the models in question and the metric we applied to quantify their effectiveness are presented in their respective publications from a technical and theoretical perspective; their original publications emphasise their technical improvements over other methods and accentuate qualities such as efficiency and ease of use. While we acknowledge the advancements and contributions of such work, according to the observations made throughout our experiments it is more accurate to regard these as proofs of concept which can deliver satisfactory results under very specific conditions and when the users have a considerable knowledge of their usage, not as a wide-ranging intuitive plug-and-play solution that can be liberally applied on any machine learning task by users that aren't familiar with the methods to their full extent. This is particularly evident with the system causability scale, which proved unclear and ill-fitting to every survey participant. Future work could benefit

from exploring models that are considered interpretable from the beginning, instead of attempting to apply explanation methods on models that aren't inherently interpretable (Rudin, 2019).

List of Figures

List of Tables

References

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Felzenszwalb, Pedro F and Daniel P Huttenlocher (2004). "Efficient graph-based image segmentation". In: *International journal of computer vision* 59.2, pp. 167–181.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

High-Level Expert Group on Artificial Intelligence (Apr. 2019). *Ethics Guidelines for Trustworthy AI*. Text. European Commission.

Holzinger, Andreas, André Carrington, and Heimo Müller (2020). "Measuring the quality of explanations: the system causability scale (SCS)". In: *KI-Künstliche Intelligenz* 34.2, pp. 193–198.

Kapishnikov, Andrei, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry (2019). "Xrai: Better attributions through regions". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4948–4957.

Lipton, Zachary C (2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Rudin, Cynthia (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019; 1: 206–15*.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Signoroni, Alberto, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, et al. (2021). "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset". In: *Medical Image Analysis* 71, p. 102046.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2013). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034*.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR, pp. 3319–3328.

Zicari, Roberto V., John Brodersen, James Brusseau, Boris Düdder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Möslein, Naveed Mushtaq, Gemma Roig, Norman Stürtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund (2021). "Z-Inspection®: A Process to Assess Trustworthy AI". In: *IEEE Transactions on Technology and Society* 2.2, pp. 83–97. DOI: 10.1109/TTS.2021.3066209.