NORDIC MACHINE INTELLIGENCE

Nordic Machine Intelligence, vol. 4, pp. 1–18, 2024
Received 15 Nov 2023 / published 19 Aug 2024
https://doi.org/10.5617/nmi.10755

Propagating Transparency: A Deep Dive into the Interpretability of Neural Networks

Ayush Somani[1,3], Alexander Horsch[1], Ajit Bopardikar[2], and Dilip K. Prasad[1]

[1] *Dept. of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway*
[2] *Department of Electrical Engineering, University of Houston, USA*
[3] *E-mail any correspondence to: ayush.somani@uit.no*

Abstract

In the rapidly evolving landscape of deep learning (DL), understanding the inner workings of neural networks remains a significant challenge. The need for transparency and accountability in DL models grows in importance as they become more prevalent in decision-making processes. Interpreting these models is key to addressing this challenge. This paper offers a comprehensive overview of interpretable methods for neural networks, particularly convolutional nets. The focus is on gradient-based propagation techniques that provide insight into the intricate mechanisms behind neural network predictions. Using a systematic review, we classify interpretability approaches that are based on gradients, dive into the theory of notable methods, and compare their strengths and weaknesses. Furthermore, we investigate different evaluation metrics for interpretable systems, often generalized under the term eXplainable Artificial Intelligence (XAI). We highlight the importance of these factors in evaluating the faithfulness, robustness, localization, complexity, randomization, and adherence to the axiomatic principles of XAI methods. Our objective is to assist researchers and practitioners in advancing towards a future for artificial intelligence that is characterized by a deeper understanding of its workings, thereby providing the desired transparency and accuracy. To this end, we offer a comprehensive summary of the latest advances in the field.

Keywords: Neural Network; Interpretability; Deep Learning; eXplainable Artificial Intelligence (XAI); Model Transparency

Introduction

The advent of highly complex Deep Learning (DL) models in recent years has revolutionized several sectors, including medical diagnostics and autonomous vehicles [1, 2, 3]. These models have consistently outperformed traditional algorithms in a variety of tasks due to their exceptional capacity to extract complex patterns from massive volumes of data. However, the same complexity that enables them also obscures their decision-making processes, making them incomprehensible to humans [4].

The lack of transparency in DL models, especially in terms of their inner workings, is often referred to as the "black-box" [5, 6]. In critical applications such as medical imaging, where decisions can have far-reaching consequences, the need for transparency and explainability becomes imperative to make informed decisions. The General Data Protection Regulation (GDPR) of the European Union also includes a provision called the "right to explanation," which grants individuals the ability to request an explanation for any automated decisions made about them [7]. This further emphasizes the critical need to explain the functioning of deep learning models.

The field of eXplainable Artificial Intelligence (XAI) has emerged in response to these challenges, with the goal of bridging the gap between a model's impressive performance with millions of parameters and the ability to explain its decision-making processes. This is achieved through a diverse set of techniques, depending on the specific model, the type of data it processes (i.e., image, text, tabular, etc.) and the specific insights required for a given application. Certain techniques prioritize the identification of key features in the input data that have the most influence on a model's decision. For instance, techniques can identify and emphasize individual pixels or areas in an image that play a significant role in a specific classification within medical imaging or autonomous systems. This tool also facilitates the analysis of textual data in reviews, social media posts, and more, and extracts valuable information on sentiment classification, including categorizing sentiment as positive, negative, or neutral. Others may emphasize the model's acquired internal logic or relationships to reach a specific decision, such as loan approval or insurance risk, in financial or risk assessment applications.

While the potential of XAI is immense, it is crucial to recognize that not all AI models possess the same level of explainability [8]. For instance, linear regression

and decision trees are typically more transparent than complex neural networks. In addition, the use of XAI varies greatly depending on the application domain and the target audience, and there is no set consensus on how to address these challenges associated with model interpretation. Domain experts may require in-depth technical explanations regarding the model's internal logic, whereas non-technical users would benefit from simpler, overarching interpretations of the model's decisions. Hence, the selection of suitable XAI techniques is closely linked to the choice of an appropriate AI model, the application and the target user ensuring the desired balance between accuracy and explainability for the specific task at hand. Recent surveys [9, 10] in XAI offer a comprehensive overview of the field's development.

Furthermore, similar to any data-driven system, AI models are susceptible to biases that exist within the data they are trained on. Uncovering these potential biases is a must for XAI techniques to guarantee that AI systems are making decisions that are ethical and fair. Considering the extensive scope of the XAI field in terms of model explainability, data explainability, feature-based techniques, and example-based techniques, this study will not attempt to cover the entire spectrum. Instead, our focus will be directed towards a particular subcategory: visual gradient-based feature attribution methods. We will dive deeper into this method, exploring its implications and progressions over time.

This paper examines the extensive range of gradient-based explanations, which offer an intuitive and effective way to examine their strengths, limitations, and various application areas. Gradient-only methods determine whether a modification in a pixel would result in a modification in the prediction. As we navigate through current advances in the field, our aim is to provide a comprehensive assessment of the state-of-the-art (SOTA) neural network interpretability, enabling researchers, practitioners, and enthusiasts to better understand and trust AI systems. In addition, we will discuss the metrics used to evaluate these explanation methods to ensure the reliability and usefulness of the explanations themselves.

Related Works

Several methods attempt to uncover the underlying reasoning of DL models. One such technique evaluates the influence of individual input features on the model's predictions using feature attribution [11, 12]. Extracting explanations that describe the input variables for images (such as pixels) used by the model to generate a prediction is a widely used strategy [13, 14]. Other studies have used feature attribution to identify sensitive input features and hidden neurons that impact predictions. Leino et al. [15] define influence-directed explanations as the average gradient of an instance measured in relation to a neuron across a range of inputs.

In general, the extensive array of pixel attribution approaches can be categorized into two types of attribution methods: (a) Occlusion- or perturbation-based methods, such as LIME [13] and SHAP [16], that leverage image manipulation techniques to generate model-agnostic explanations, and (b) Gradient-based methods involve calculating the gradient of the prediction or classification score in relation to the input features. The various gradient-based methods primarily vary in their approach to computing the gradient. Here, both attribution based approaches have a common characteristic that the explanation has the same dimensions as the input image (or can be effectively projected onto it). In addition, they assign a numerical value to each individual pixel. This can be interpreted as the significance of the pixel in relation to the prediction or classification of the image.

Another useful categorization for pixel attribution methods is based on the fundamental theory that determines whether a change in a pixel would lead to a change in the prediction. Two popular examples are the vanilla gradient [11] and grad-CAM [17]. The attribution based solely on the gradient can be interpreted in the following manner: If the values of a pixel (representing features such as color or intensity) were to increase, the predicted class probability would correspondingly increase (for positive gradient) or decrease (for negative gradient). The magnitude of the gradient provides an indication of the relative sensitivity of the prediction to changes in that pixel's value. However, the exact impact on the prediction also depends on additional variables such as the model's architecture, the specific feature being modified, and the interactions between different features.

Path-attribution methods involve comparing the current image with a reference image, which may be an artificial "zero" image, such as a completely gray image. The disparity between the actual prediction and the baseline prediction is attributed to individual pixels. Note that the selection of the reference image (distribution) significantly impacts the explanation. The common practice is to utilize a "neutral" image. The category includes model-specific techniques such as Deep Taylor Decomposition and Integrated Gradient [18], as well as model-agnostic techniques such as LIME and SHAP. Certain path-attribution methods are considered "complete" as they calculate the difference between the prediction of an image and the prediction of a reference image by summing up the relevance scores of all input features.

Gradient-based propagation methods are noteworthy due to their ability to provide a high level of granularity, versatility, and direct correlation with input features. They have a deep understanding of the flow of information through networks, which makes their insights particularly valuable across various sectors. Methods such as saliency maps pinpointed influential pixels in input images [19], and techniques such as the vanilla gradient method identified pixels that have the most influence on the model's output score [11]. Over time, various methods such as guided

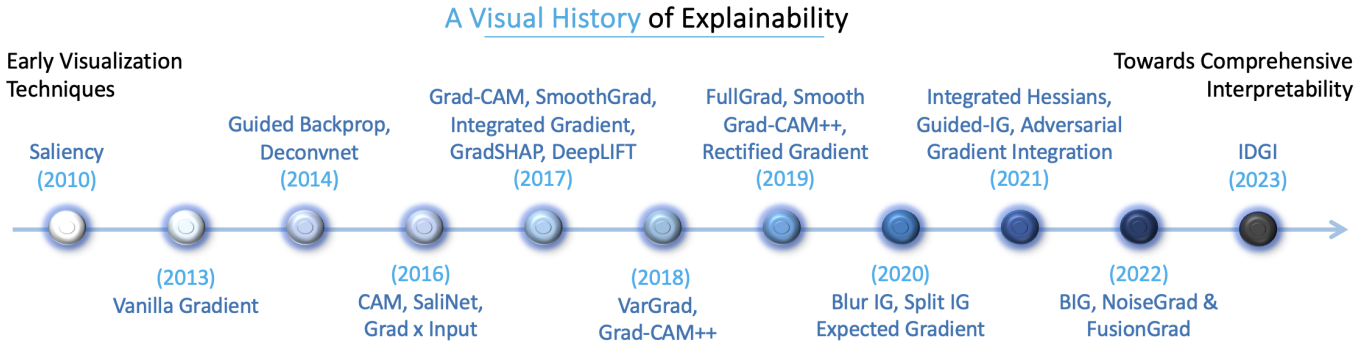

Figure 1: Evolution of propagation-based visualization methods for explainability.

backpropagation and grad-CAM have improved the field by providing clearer and more generalized insights [17, 14]. Although gradient-based feature attributions [20, 21], continue to be a subject of interest, some concerns have been raised about their reliability, as highlighted in recent research [22, 23]. One of these challenges is the vulnerability of these methods to input noise, which can affect their accuracy. Furthermore, there is concern that these methods may sometimes provide misleading explanations[24]. This is an area of active investigation [18, 8].

Recent advances emphasize the need to perform sanity checks to assess the credibility of interpretations [25]. Several metrics have been introduced that assess the axiomatic principles, implementation invariance, and consistency of interpretability methods [26, 27, 24]. Quantus and OpenXAI are frameworks that have been developed to enhance the dependability of interpretability estimators [28, 29]. While many methods are largely model-agnostic, there is a growing niche that focuses on model-specific interpretability, tailoring self-explainable frameworks to particular architectures.

In this study, we comprehensively investigate the application of backpropagation methods and their potential for XAI research. When sourcing our insights, we meticulously reviewed the seminal work shown in Figure 1, taking into account both the foundational literature and the most recent research from reputable conferences and journals. This approach ensures a comprehensive understanding of the development and possibilities of gradient-based interpretability in neural networks without favoring any specific approach. However, we exclude the discussion of more expensive visual explanation techniques that employ a local or global surrogate model, such as LIME [13] and SHAP [16] respectively.

Gradient-based Visual Interpretation

This section outlines different techniques that use gradient information in post-hoc model settings and build attribution maps for visual explanations. These methods make use of the gradient of the network output (logits or soft-max probabilities) with respect to (w.r.t.) the input features. Consider a network with an N-dimensional input $x = \{x_i\}_{i=1}^{N} \in \mathbf{R}^N$ and a C-dimensional output $S(x) = \{S_c\}_{c=1}^{C} \in \mathbf{R}^C$. In this context, C is the total number of classes, S_c can be a class score (logit) or soft-max probability and $S_c(x)$ represents the network's score function. The "gradient" term $(\delta/\delta x\ S_c(x))$ estimates the attribution map, $A^c = \{A_i^c\}_{i=1}^{N} \in \mathcal{R}^N$ which captures the importance of each input feature for a specific output class c. In computer vision tasks, we consider Convolutional Neural Networks (CNNs) with image input, i.e., the pixels of the image are considered input features.

The idea of attention maps operates on differentiable models and leverages the network's gradient acquired through forward-backward propagation to capture the relationship between input and output (ref. Figure 2). This is achieved by identifying influential regions or sequences in the input data [42]. By analyzing the relevance of each feature and its impact on a model's prediction, we not only gain valuable insights but also uncover biases, identify anomalies, or flag excessive dependencies on specific inputs, thus strengthening trust and guiding model refinement [43]. The uniqueness is attributed to:

- Direct correlation with input features: Unlike model-agnostic techniques that offer a bird's-eye view, such as in LIME [13] where an explanation may be derived locally from the records generated randomly in the neighborhood of the target to be explained, gradient-based methods directly correlate explanations with specific input features. This direct correlation allows for more actionable insights to be gained from the explanations [44].
- Flexibility & scalability: These methods are inherently adaptable, seamlessly fitting diverse DL architectures without significant modifications [3].
- High fidelity in medical applications: In medical imaging, where interpretability can have life-altering implications, certain methods like Grad-CAM and integrated gradients have proven to be effective in generating explanation maps that closely match expert annotations in various tasks, such as tumor detection in radiology images [45].

Table 1: Detailed comparison of gradient-based methods for visual interpretability

Method	Year	Application	Evaluation Method	Interpretability Impact
Saliency [19]	2010	Direct gradients w.r.t. input	Highlights influential pixels in images	Sensitive to input noise
Vanilla Gradient [11]	2013	Direct gradients w.r.t. input	Pinpoints pixels that most affect the output score	Can produce noisy explanations
Guided Backprop [17]	2014	Combines positive gradients with ReLU activations	Emphasizes pixels that positively influence the final prediction	May not be suitable for all architectures
Grad x Input [30]	2016	Element-wise multiplication of input and its gradient	Offers pixel-wise decomposition of the output	May produce artifacts
Grad-CAM [14]	2017	Weighted combination of feature maps & gradients	Visualizes regions in images that activate specific feature maps	Not always precise in localization
Integrated Gradients [18]	2017	Path integration of gradients	Decomposes prediction output over input features	Computationally intensive
GradientSHAP [16]	2017	Integrating concepts from both Integrated Gradients and SHapley Additive exPlanations (SHAP).	Computes attributions by averaging gradients over multiple background samples.	Provides stable, averaged explanations, effectively highlighting semantically meaningful input regions.
DeepLIFT [31]	2017	Decomposes output prediction to contributions of all neurons to the input features	Differentiates feature activations from reference activations and assigns contributions.	Provides high-resolution attributions that can distinguish contributions of different features
SmoothGrad [12]	2017	Averages noisy versions of input gradients	Reduces noise in gradient-based explanations	Requires multiple evaluations
VarGrad [32]	2018	Incorporates variance of gradients	Requires multiple gradient evaluations	Enhances saliency by considering gradient variance
FullGrad [33]	2019	Computes the gradients of the biases from all over the network, and then sums them	Response decomposition into input sensitivity & per-neuron sensitivity components	Satisfy both completeness & weak dependence properties of saliency maps
Expected Gradient [34]	2020	Averages gradients over possible inputs	Provides a more stable and averaged explanation	Computationally intensive
Blur IG [35]	2020	Averages gradients over blurred versions of the input	Reduces noise and offers smoother explanations	Introduces blurring artifacts
Integrated Hessians [36]	2021	Extends IG by incorporating second-order derivatives	Provides deeper insights into model sensitivity	Improves interpretability by considering decision boundary curvature
Guided Integrated Gradients [37]	2021	Combines Guided Backprop with IG for refined attributions	Precisely highlights critical paths than either method alone	Offers detailed and specific feature attributions
Adversarial Gradient Integration [38]	2021	Uses adversarial examples to weigh gradients	Identifies model vulnerabilities and robust features simultaneously	Provides insights into model weaknesses and robustness
Boundary-based IG [39]	2022	Understands model predictions via decision boundary dynamics	Highlights how input perturbations shift the decision boundary	Clarifies model decisions by visualizing boundary changes
NoiseGrad and FusionGrad [40]	2022	Incorporates noise in the gradient computation	Offers diverse explanations by considering gradient noise	Can be sensitive to noise type
Important Direction GI [41]	2023	Targets the most influential directions in input space	Isolates and illustrates key input directions affecting predictions	Offers a focused analysis of features most critical to model decisions

We will examine the details of each gradient-based method, as it offers a comprehensive understanding of the overall approach that many other methods follow, summarized in Table 1.

1. Saliency (2010). Saliency maps are designed to generate a heatmap overlay on an input image [19]. The map computes the gradient $S(x) = ||\nabla_x y||$ of the output prediction y in relation to the input image x. The magnitude of the gradient for each pixel indicates its influence on the prediction. Note that this is a foundational method and often produce noisy results. Variations to enhance clarity of saliency maps include SmoothGrad [12].

2. Vanilla Gradient (2013). The method [11] is one of the earliest pixel attribution approaches. It calculates the basic gradient of the loss function $S_{grad}(x) = \nabla_x s_c(x)$ for the output score $s_c(x)$ of a specific target class c w.r.t. the input image x. It then assigns a value of zero to all other classes. This gives us a map of the size of the input features that either show the absolute values or highlight negative and positive contributions separately. This helps identify pixels that would cause the most significant change in the output score when altered. This is achieved using gradient ascent to iteratively modify the input image to maximize the logit outputs [46] shown in Figure 2. Avanti et al. [31] discuss the utility of vanilla gradients for model interpretation and highlight their potential limitations, such as saturation problems and noisy visualizations.

The method serves as the basis for more sophisticated

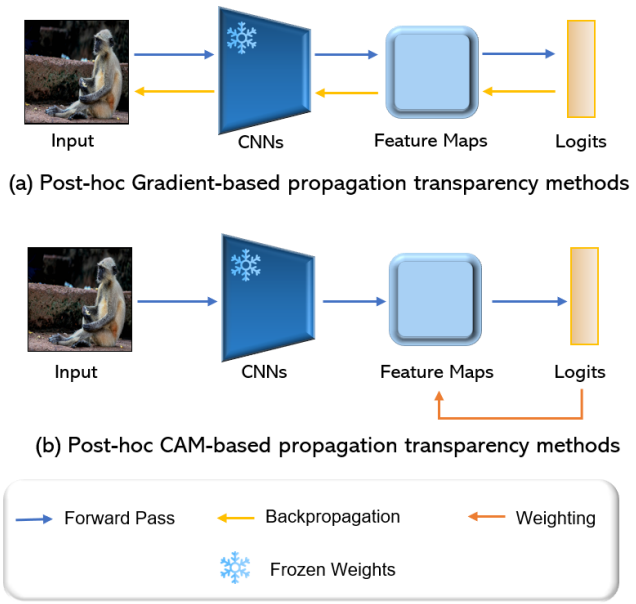

Figure 2: General flowchart of propagation-based methods for visualizing the relationship between input and output by leveraging the differentiable neural network's gradient obtained from one or more forward and backward passes. (a) Gradient-based methods tends to backpropagate gradients from logits to input space. (b) Existing CAM-based methods focus on generating better weighting schemes to obtain weighted-sum attention maps.

techniques and domains, where they might outperform more advanced gradient-based methods. In order to better understand, when an image x is passed through a CNN, it gives a score $s_c(x)$ for the corresponding class c. The score depends on the input image in a complex and non-linear way. The rationale for using the gradient is to estimate the score by applying a first-order Taylor expansion in equation 1 below.

$$\begin{aligned} s_c(x) &\approx w^T x + b \\ w &= \frac{\delta s_c}{\delta x}|x_0 \end{aligned} \tag{1}$$

where, w represents the derivative of the score. However, there is ambiguity surrounding the implementation of the backward pass for gradient calculations when working with nonlinear units such as ReLU (Rectifying Linear Unit) that "remove" the sign. Here, the input to layer f^{l+1} is the output of the ReLU function which is defined as $\text{ReLU}(f_i^l) = max(0, f_i^l)$ from the previous layer f^l. When performing a backpass, it can be difficult to decide whether to assign a positive or negative activation. When a neuron's activation is zero, it becomes uncertain which value should be propagated backwards. When it comes to vanilla gradient, the ambiguity is resolved in the following manner:

$$\frac{\delta f}{\delta f_i^l} = \frac{\delta f}{\delta f_i^{l+1}} \cdot \mathbf{I}(f_i^l > 0) \tag{2}$$

with f being the final output of the model. Here, the element-wise indicator function $\mathbf{I}$ is used to assign a zero value to negative activations at lower layers and a value of one to positive or zero activations.The method backpropagates the gradient up to layer $l+1$, and then simply nullifies the gradients where the activation at the layer below is negative. However, if the activation falls below zero, ReLU caps it at zero and remains unchanged thereafter, leading to the problem of activation saturation.

3. Guided Backprop (2014). Introduced as an improved visualization technique for DNNs, guided backpropagation [17] modifies the standard backpropagation, suppressing negative gradients to zero for clearer and sharper results compared to using the vanilla gradients. The idea is that negative gradients in standard backpropagation can occasionally suppress evidence, making visualizations noisy. By focusing only on positive gradients, guided backprop aims to highlight areas in the input image that positively contribute to the model's decision. This sharpens the focus on the most influential input regions in the input image for the model's prediction.

Mathematically, the activation for the forward pass through multiple layers l to generate a feature map for the final convolutional layer is defined as $f_i^{l+1} = \text{ReLU}(f_i^l) = max(0, f_i^l)$, where f^0 corresponds to the input image x and f^{out} corresponds to the final output. The typical process of standard backpropagation through the l layers to generate the reconstructed image R^0 is defined as follows:

$$R_i^l = (f_i^l > 0) \cdot R_i^{l+1} \quad \text{where, } R_i^{l+1} = \frac{\delta f^{\text{out}}}{\delta f_i^{l+1}} \tag{3}$$

Equation 4 defines the guided backpropagation that suppresses negative values during the backward pass.

$$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \tag{4}$$

4. Grad x Input (2016). A straightforward and computationally efficient method for generating an explanation for CNN's decision. The method [30] involves element-wise multiplication of the input image by its gradient to produce heat map $S(x)_{GxI} = x \odot \nabla_x F(x)$ that is more localized and focused compared to previous gradient-based methods. This makes it easier to identify the pixels in the input image that contribute most to the model's output.

5. Grad-CAM (2017). Grad-CAM [14], short for gradient-weighted class activation mapping, makes class-discriminative visual explanations for decisions made by a wide range of CNN architectures. Grad-CAM generates a coarse localization map by propagating the gradients of the target class into the final convolutional layer. In simpler words, it aims to obtain a deeper understanding of

the specific areas of an image that a convolutional layer prioritizes when making a particular classification. This decision of interest can be the class prediction (which we find in the output layer), but it can theoretically be any other layer in the neural network.

Let us analyze Grad-CAM from a practical standpoint. The first convolutional layer of a CNN receives images as input and generates feature maps that encode learned features. The higher-level convolutional layers perform a similar function, but they receive the feature maps from the preceding convolutional layers as input. For the initial approach, we can visualize the raw values of each feature map, calculate the average across the feature maps, and then overlay this information onto the image. This would not be useful as the feature maps encode information for all classes, whereas our focus is on a specific class. In order to properly calculate the average over the feature maps, it is essential to assign a weight to each pixel based on its gradient.

Let y^c be the class score, A^k be the feature map of the last convolutional layer, and α_K^c be the weights that capture the importance of the feature map k for the target class c. Then the global average pooling of the gradient maps is computed as follows:

$$\alpha_K^c = \frac{1}{z} \sum_i \sum_j \frac{\delta y^c}{\delta A_{i,j}^k} \tag{5}$$

Here, z typically represents the normalization factor, which is the total number of elements (pixels) in the feature map A^k. Each feature map "pixel" is weighted by the gradient of the class. Indices i and j correspond to the width and height dimensions. Finally, the localization map is defined as the linear combination followed by ReLU, as shown in equation 6.

$$L_{\text{Grad-CAM}^c} \in \mathcal{R}^{u \times v} = ReLU\left(\sum_k \alpha_k^c A^k \right) \tag{6}$$

The resultant heatmap identifies areas that have an impact, either positive or negative, on the target class. This heatmap is passed through the ReLU function, similar to what we learned for the vanilla gradient, focusing solely on the regions that have an impact on the target class. It is important to note that the word 'pixel' might be misleading here, as the feature map is actually smaller than the image due to the presence of pooling units. However, it is then mapped back to the original image after scaling the explanation map to the interval [0,1] and superimposing it on the original image for visualization. The method is class-discriminative, which means it can highlight different regions for different target classes.

Selvaraju et al. [14] provide empirical results, showing the utility of Grad-CAM in various tasks such as image classification, captioning, and visual question answering. It also demonstrates that Grad-CAM visualizations can be used for weakly supervised object localization. In the following year, building on the previous work, Chattopadhay et al. [47] presented Grad-CAM++ as a generalization of Grad-CAM. The method provided better visualization using only the positive partial derivatives of the last convolutional layer. Recently, a study [48] focused on extending existing Grad-CAM techniques to the 3D domain using the Inflated Inception 3D pipeline for video-based Human Action Recognition (HAR). The study shows enhanced understanding of spatio-temporal information in HAR models, providing valuable insights into the decision-making process of video-based AI systems.

6. Integrated Gradients (2017). Gradients represent infinitesimal prediction variations caused by infinitesimal feature changes. So, when an input x leads to a high prediction (e.g., a high probability for a particular class), just looking at small changes in prediction (using standard gradients) may not provide a clear understanding of why the model made that prediction. This is because the relationship between input features and the prediction might be complex and non-linear. Furthermore, vanilla gradients are subject to the saturation problem [18], in which the gradients of certain features are close to zero despite the fact that the model substantially relies on those features. When using Integrated Gradients (IG) [18] to explain something, gradients are typically accumulated along the way from a baseline input to the actual input. It provides a way to decompose the difference in output prediction (between the baseline and actual input) into contributions from each feature. The method satisfies two important axioms: (i) sensitivity (detecting features that make a difference) and, (ii) implementation invariance (consistent attributions across functionally equivalent networks).

The assignment of importance score to each input feature is computed by defining a baseline input (e.g., a black image for an image network or a zero embedding vector for a text model) and linearly interpolating between the baseline input $\bar{x}$ and the actual input image, x. Equation 7 computes the gradient for the i^{th} feature along this interpolation path.

$$IG_i(x) ::= (x_i - \bar{x}_i) \times \int_{\alpha=0}^{1} \frac{\partial S_c(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i} d\alpha \tag{7}$$

here, $\partial S_c(x)/\partial x_i$ is the gradient of $S_c(x)$ along the i^{th} dimension. Integrating over a path prevents local gradients from becoming saturated. If a straight line is established, IG maintains symmetry. Note that this explanation can be computationally expensive in comparison to some simpler methods. In 2021, Hesse et al. [49] demonstrated that given a non-negatively homogeneous model, IG with a zero baseline is equivalent to Input × Gradient.

7. GradientSHAP (2017): A method that combines ideas from Integrated Gradients and SHapley Additive exPlanations (SHAP) to produce feature attributions for deep learning models. GradientSHAP [16] computes attributions by averaging gradients over multiple background samples and then sums up the difference between the model's output for the input and its expected output for the reference. Thereby, the resulting attributions possess desired properties, such as completeness and symmetry preservation. While 'completeness' ensures all parts of the prediction are explained by the attributions,'symmetry preservation' aims to ensure equal contributions from features are treated equally in attributions. Details on different desired properties and their importance in ensuring effective model interpretability are discussed in the following subsection. Recently, López et al. [50] used GradientSHAP in conjunction with other explainable attribution methods to identify two subgroups of Pancreatic Ductal Adenocarcinoma (PDAC) patients with differing prognoses and biological factors, highlighting the role of DNA methylation. This improved understanding of how the model learned hidden features from multi-omics data to make critical healthcare predictions.

9. DeepLIFT (2017). Deep Learning Important Features (DeepLIFT) [31] dissects a network's output prediction on a specific input by backpropagating the contributions of all neurons in the network to each feature of the input, similar to Layerwise Relevance Propagation (LRP) [51]. Each unit i is assigned an attribution $r_i^{(L)}$ that compares the relative effect of each neuron's activation at the original network x to its reference activation $\bar{x}$ and ranks contributions based on the difference (shown in equations 8 & 9). To calculate the reference values $\bar{z}_{ji}$ for all hidden units, a forward pass through the network is performed, with the baseline $\bar{x}$ as input. Each unit's activation is recorded.

$$r_i^{(L)} = \begin{cases} S_i(x) - S_i(\bar{x}) & \text{if } i \text{ is the target unit} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$r_i^{(l)} = \sum_j \frac{z_{ji} - \bar{z}_{ji}}{\sum_{i'} z_{ji} - \sum_{i'} \bar{z}_{ji}} r_j^{(l+1)} \tag{9}$$

where, $r_i^{(L)}$ and $r_i^{(l)}$ represent the relevance scores in the final layer L and at any layer l, respectively. The attributions at the input layer are defined as $R_i^c(x) = r_i^{(1)}$. The weighted activation $\bar{z}_{ji} = w_{ji}^{(l+1,l)} \bar{x}_i^{(l)}$ of a neuron i onto neuron j when the baseline $\bar{x}$ is fed to the network. Equation 9 describes the "Rescale rule" that was used in the original formulation of the method. The other rule, "Reveal-Cancel" was also proposed in the paper [31].

DeepLIFT can identify dependencies that other techniques miss by optionally taking positive and negative contributions into account separately. Scores can be efficiently computed with a single backward pass. Compared to the attribution defined for IG in equation 7, the attribution for the DeepLIFT can be defined as follows:

$$\begin{aligned} \text{DeepLIFT}_i(x) &= (x_i - \bar{x}_i) \cdot \frac{\partial^g S_c(x)}{\partial x_i}, \\ \text{where, } g &= \frac{f(z) - f(\bar{z}}{z - \bar{z}} \end{aligned} \tag{10}$$

It is important to acknowledge that while Shrikumar et al.'s [30] Grad x Input offers a computationally efficient way to generate explanation maps, DeepLIFT [31] provides a more comprehensive and theoretically grounded approach to feature attribution.

8. SmoothGrad (2017). Smilkov et al. [12] found that the gradients (or derivatives) of the model's prediction w.r.t. the input can change sharply around the input image. This happens even though the changes in the images are so subtle that they are visually indistinguishable to human eyes. The authors emphasize that noise may appear due to high local variations of gradients. SmoothGrad works by adding noise to the input multiple times, computing the gradient each time, and then averaging the results. This process smoothens and sharpens the visualization, mitigating issues like noisy gradients and providing clearer insights into the most influential part of the input image for the model's decision. Empirical evaluations demonstrate that SmoothGrad produces more visually coherent and interpretable saliency maps compared to standard gradient visualizations [52].

For an input image x, the smoothed explanation $E_{SGrad}(x)$ is calculated in equation 11 using σ, the standard deviation of Gaussian noise.

$$E(x)_{SGrad} \approx \frac{1}{N} \sum_{i=1}^{N} E(x + \mathcal{N}(0, \sigma^2)) \tag{11}$$

where, N be the number of times the noise is sampled, and $\delta_i \sim \mathcal{N}(0, \sigma^2)$ represents Gaussian noise with zero mean and variance σ^2. Although it is simpler to implement, the number of noise samples (N) and the noise level (σ) are hyperparameters that might require adjustment.

10. VarGrad (2018). SmoothGrad [12] was designed with the objective of providing more stable and consistent explanation maps compared to other gradient-based methods. While effective, it still requires multiple computations at the cost of efficiency. In the following year, VarGrad [32] aims to achieve similar results with less computational overhead. VarGrad introduces a variance-based weighting to gradients, which helps to emphasize the most influential features while suppressing noise. The underlying idea is that pixels with consistently high gradient variance across different perturbations of the input are more likely to be truly influential. By focusing on regions with high gradient variance, VarGrad aims to

identify areas in the input that are consistently influential across multiple perturbations. Nevertheless, the choice of perturbation type and magnitude could influence the results.

$$E_{VGrad}(x) \approx \frac{1}{N}\sum_{i=1}^{N}[E(x+\delta_i) - E_{SGrad}(x)]^2 \qquad (12)$$

Although both denoising strategies can improve gradient-based explanations, Seo et al. [53] conclude that SmoothGrad does not smooth the gradient of the prediction function, whereas VarGrad captures higher-order partial derivatives rather than being dependent on the gradient of the prediction function.

11. FullGrad (2019). Existing visual explanation methods often lack the ability to capture both fine-grained detail and a holistic view of how a model uses structural relationships within the input image. FullGrad [33] provides a novel approach to decompose the output prediction into input gradients and bias gradients simultaneously, particularly for convolutional networks.

Mathematically, for a ReLU neural network f with biases $b \in R^F$, equation 13 adds the gradient w.r.t. the input x, which is given by $\nabla_x f(x;b)$, and the gradient w.r.t. the bias term, denoted by $\nabla_b f(x;b)$ which includes the latent biases in batch normalization layers, as well as explicit biases in convolutional and fully connected layers.

$$f(x;b) = \nabla_x f(x;b)^T x + \nabla_b f(x;b)^T b \qquad (13)$$

In a CNN, the bias parameters have the same spatial structure as the feature map because of weight sharing and sliding window mechanisms used during convolution. After post-processing, we can visualize the bias contribution $\nabla_b f(x;b)^T b$ of the explanation map in equation 14.

$$S(x) = \psi(\nabla_x f(x;b)^T x) + \sum_{l\in L}\sum_{c\in c_l}\psi([\nabla_b f(x;b)^T b]_c^l) \qquad (14)$$

where, L denotes the total number of convolutional layers. In each layer l, c_l represents the number of channels. The function $\psi(.)$ includes various post-processing steps, such as upsampling, abstracting, and applying Min-Max normalization. Upsampling $\nabla_b f(x;b)^T b$ is necessary to counteract the downsampling from the forward pass and to ensure the gradient dimensions align with the image dimensions at each layer l.

The method is distinguished by its comprehensive decomposition of the neural network response into two components: (i) input sensitivity, reflecting the importance of individual input pixels, and, (ii) neuron sensitivity, accounting for the importance of groups of pixels and their structure. This provides a comprehensive understanding of both local and global contributions to the network's decisions compared to traditional methods, producing sharper and more object-region confined saliency maps. Thereby, addresses the limitations of traditional explanation map-based interpretability methods by satisfying both completeness and weak dependence properties, which are typically not simultaneously achievable in other methods. However, its practical use deviates because the bias gradient is ignored in fully connected layers and the upsampling steps are included. In addition, the application of FullGrad to fields like tabular data and natural language processing (NLP) has yet to be explored.

12. Expected Gradient (2020). The method [34] builds on the concept of IGs but incorporates expectations over multiple perturbed inputs. By default, IG uses the all-zero vector as its baseline. However, when the target object's body is black, IG fails to effectively highlight that area. Alternative baselines for IG includes an image with the maximum distance from the current input, a Gaussian-blurred image, an image with random pixel values, and a black-and-white baseline [52]. However, each baseline option has its advantages and disadvantages. A common approach is to average the attribution scores from various baselines drawn from a distribution.

Expected Gradient averages out noise and highlights consistently influential features. By integrating over a distribution of perturbed inputs with Riemann integration, expected gradient offers a more holistic view of interpretable and consistent gradient-based visualizations. The method is particularly effective at highlighting semantically meaningful regions in the input.

13. Blur Integrated Gradient (2020). A variant of the IG method, seeks to explain the prediction by combining gradients from both the frequency and spatial domains. Blur Integrated Gradient (Blur-IG) [35] works by successively blurring the input image with a Gaussian blur filter and integrating over a straight-line path in the input space between the blurred image and the original image. The blurring process helps suppress noisy or irrelevant features that may arise during interpolation when new features are introduced along the integral line, emphasizing only the most influential regions in the input.

14. Integrated Hessians (2021): Although IGs offer reliable attributions, calculating them can be computationally expensive, especially for high-dimensional inputs. Integrated Hessians [36] aim to approximate second-order derivative information explaining pairwise interactions between features within neural networks with fewer computations. This method integrates the second derivatives (Hessians) of the model's output along a path from a baseline $\bar{x}$ to the input x, shown in equation 15. This typically involves approximations and sampling techniques to make computations of feature attributions feasible and potentially faster compared to standard IGs. This makes it particularly valuable for analyzing large models and/or

high-dimensional data.

$$IH_i(x) = (x_i - \bar{x}_i) \times \int_{\alpha=0}^{1} \frac{\partial^2 S_c(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i \partial x_j} d\alpha \quad (15)$$

The method has demonstrated superior efficiency and effectiveness in capturing feature interactions compared to existing methods, particularly when the number of features is large. Integrated Hessians provides a more nuanced understanding of how feature interactions influence model output, extending beyond specific types of neural networks.

15. Adversarial Gradient Integration (2021): Conventional gradient-based methods might struggle with the ambiguity and inconsistency in model interpretations when the input-output relationship is locally non-smooth. The rationale behind selecting an uninformative baseline in IG is not adequately justified for certain tasks [52]. Adversarial Gradient Integration (AGI) [38] introduces a baseline-free method. AGI integrates gradients for the model f from adversarial examples belonging to a different class j (different from the current prediction) to the target example i along the steepest descent path to calculate the reliable importance in these scenarios. Equation 16 presents the aggregation of gradients in the descending steps, where "til adv" implies that the process stops either when an adversarial example is found or when the maximum predefined step limit is reached. The number of non-target classes may be large, and then AGI calculates the attribution score by averaging it over multiple randomly selected classess.

$$AGI_i(x) = \int_{\text{til adv}} -\nabla_{x_i} f(x) . \frac{\nabla_{x_i} f_j(x)}{|\nabla_{x_i} f_j(x)|} d\alpha \quad (16)$$

This method is distinct in that it does not does not rely on a predefined reference point, which helps avoid inconsistencies that arise from arbitrary reference selections often used in other gradient-based methods. The method demonstrated improved consistency and clarity in assigning input feature contributions to model predictions, surpassing existing methods. It showed proficiency in handling cases with a reliable explanation framework by eliminating the need for reference input.

16. Boundary-based Integrated Gradients (2022): The calculation of path integrals in traditional IGs can be sensitive to the choice of baseline and path. Boundary-based IG [39] aims to reduce this sensitivity by targeting decision boundaries and providing more precise and meaningful attributions. The method utilizes a gradient ascent approach, where gradients are calculated not just with respect to the input features but also considering their movement towards or away from the decision boundary. This involves iteratively adjusting the input x, assuming it lies within a polytope P based on the gradient information as shown in Equation 17 to probe the decision boundary's influence on the prediction.

$$BIG_i(x) = IG_i(x, x') \quad (17)$$

If the model can be represented as a local linear model $f(x) = w_p x + b_p$, the gradient w_p may not accurately account for the model's decision if the polytope is far from the decision boundary. Therefore, Wang et al. [39] introduce the closest adversarial example x' of x, where $Vx_m, ||x_m - x|| < ||x' - x|| \rightarrow F(x) = F(x_m)$ to facilitate the explanation of $f(x)$. x_m refers to a point within the polytope P that is closer to the decision boundary than the original input x. As a result, x_m is a more relevant point for assessing the impact of the decision boundary on the prediction made by the model F. Equation 17 is defined by substituting the baseline of Integrated Gradients (IG) with the adversarial example and then aggregating the gradients in a manner similar to IG.

The method offers a more nuanced understanding of feature influence, particularly in cases where small changes in inputs near decision boundaries can lead to significant changes in output. It is worth mentioning that although the attributions are potentially more robust to the choice of baseline and integration path compared to standard IGs, the method's reliance on accurate identification of decision boundaries can be challenging in complex models or datasets where boundaries are not well-defined.

17. NoiseGrad & FusionGrad (2022): Bykov et al. [40] introduce two gradient-based interpretation methods: NoiseGrad and FusionGrad. NoiseGrad focuses on assessing the robustness of gradient-based explanation maps by introducing controlled noise into the input data. The objective is to gauge how stable and reliable a saliency map by introducing perturbations to the input. FusionGrad combines multiple gradient-based interpretation methods to produce a composite explanation map, with the goal of leveraging the advantages and minimizing the limitations of each individual method.

The empirical evaluations in [40] presents that both NoiseGrad and FusionGrad produce more interpretable and consistent visualizations than some existing methods. The techniques effectively highlight influential regions in the input while suppressing noise and artifacts.

18. Important Direction Gradient Integration (2023): A common challenge in IG-based methods and their variants is that, despite being SOTA in interpreting deep neural networks, they often integrate noise into their explanation maps. Yang et al. [41] highlight that one reason is the multiplication of the gradient and the path segment during each step of Riemann integration. The Important Direction Gradient Integration (IDGI) aims to address this issue by identifying the most relevant direction of gradient change for each input feature. IDGI operates by

decomposing each gradient vector into two components: the 'important direction,' which contributes to changes in the model's prediction, and the 'noise direction,' which does not. The method then integrates only the important direction components across the input space to emphasize changes that affect the output prediction significantly.

Extensive testing showed that IDGI significantly improved interpretability metrics by effectively reducing the noise involved in traditional IG computations. This improvement was consistent across different datasets and model architectures, confirming the efficacy of focusing on important directions in gradient-based explanations. IDGI adheres to key axioms of explanation methods such as completeness and sensitivity, indicating robust theoretical foundations. However, depending on the specific underlying IG method used, it is possible that it does not satisfy the linearity axiom.

In addition to these methods, other XAI techniques, such as model simplification, surrogate models, knowledge distillation, and rule extraction, also play a crucial role in interpreting neural networks [13, 54, 51, 55]. These techniques are outside the scope of the paper.

Evaluation Metrics for XAI

While interpretability is inherently subjective, the quality of the explanations can be objectively assessed on the basis of simplicity, consistency, effectiveness and completeness for their practical use. With the field's evolution, several metrics have emerged to assess and validate interpretability techniques. These evaluation metrics can be compartmentalized into six domains, summarized in Table 2.

When two explanations differ greatly from each other, the obvious question is - Which of them is the most accurate? One notable evaluation technique is pixel flipping [51], which evaluates the precision of an explanation by examining how the omission of identified influential characteristics affects the prediction. However, this method requires an explicit feature removal scheme tailored to the type of data and prediction task.

Presently, it is evident that relying solely on visual assessments is inadequate for determining the optimal strategy to compare and explain different interpretable techniques [28]. The evaluation should scrutinize AI explainability methods against several benchmarks, encompassing consistency, stability, and fidelity to the model's decisions. In this paper, we provide an overview of the six evaluation criteria used to assess the effectiveness of interpretability methods:

1. Faithfulness. The metrics assess the degree to which an explanation accurately reflects the actual behavior and decisions of the model. For example, the degree to which the explanation satisfies recognized clinical features or expert annotations in medical imaging can serve as a measure of faithfulness [56].

The domain of 'faithfulness' in model interpretability deals with the intricate challenge of ensuring that explanations provided by models are genuinely representative of their underlying decision-making processes. The work of Bhatt et al. [26] sheds new light on the subject by looking at feature-based model explanations. The authors propose a framework that allows for the quantification and comprehension of the impact of each input feature on the model's output. Meanwhile, Dasgupta et al. [62] extensively explore the concept of faithfulness in 2022, advocating the importance of consistency and sufficiency in explanations. The seminal work on pixel-wise explanations by Bach et al. [51] uses Layer-wise Relevance Propagation (LRP) to uncover the nuances of classifier decisions, providing additional evidence in favor of this perspective. Arya et al. [57], among others, have conducted critical reviews of many different local feature attribution methods. Table 3 points to the importance of having strong and consistent evaluation methodologies.

However, the journey of faithfulness does not saturate here. Samek and his team's research on the visualization of DNN learning [58], emphasizes the importance of transparency in the complex and intricate nature of these networks. Additionally, Ancona et al.'s [70] description of the crucial role of gradient-based attribution methods emphasizes the pressing need for a comprehensive understanding of these methods. The discourse on faithfulness is enriched by significant contributions, such as the Iterative Removal Of Features (IROF) evaluation metric [71] and the focus on infidelity [60]. They highlight both the difficulties and opportunities of ensuring that model explanations are authentic, clear, and, above all, accurate in reflecting their functioning.

2. Robustness. It is critical to not only generate explanations but also to ensure their resilience, emphasizing that similar inputs should inherently produce analogous explanations [59]. The metric evaluates the stability of explanations in the presence of minor perturbations or changes to the input. This is particularly crucial in fields such as radiology, where minor differences in imaging can result in markedly distinct diagnoses [72]. Yeh et al.'s [60] study emphasizes the significance of objectively evaluating explanations, particularly those based on saliency. Meanwhile, Montavon et al. [61] offer comprehensive methods that emphasize continuity in explanations, underscoring the importance of consistency. These theme of faithfulness and trustworthiness in explanations resonates in the works of Agarwal et al. [63], who delve into the stability of explanations, both in terms of input and representation. Together, these studies shown in Table 4 collectively emphasize an important point: while interpretability is crucial, the robustness of the resulting explanations are equally, if not more, essential.

3. Localization: The metric focus on the precision of

Table 2: Metrics and their major characteristics

Metric	Definition & Purpose	Categories	Refer
Faithfulness	Measures alignment of explanation with model's actual behavior	Quantitative	Table 3
Robustness	Assesses stability of explanations under minor input perturbations	Quantitative, Visual	Table 4
Localization	Gauges precision of explanation in identifying influential input regions/features	Quantitative, Visual	Table 5
Complexity	Measures simplicity and comprehensibility of the generated explanation	Qualitative	Table 6
Randomization	Evaluates explanation reliability under model weight randomization	Quantitative	Table 7
Axiomatic Metrics	Uses theoretical principles like implementation invariance & sensitivity for evaluation	Quantitative	Table 8

Table 3: Summarizing the relevance of the five properties within the framework of evaluating faithfulness

Property	**Key Contribution & Assumptions**	**Influence & Impact**
Faithfulness Correlation [26]	Evaluates how each input attribute affects model output for a specific data point. Emphasizes the importance of quantitative evaluation criteria.	Serves as a foundational metric in explainable AI, ensuring interpretations align with model decision-making mechanics.
Faithfulness Estimate [56]	Focuses on a posteriori explanations for trained DL models around specific predictions to ensure they reflect model's actual dynamics.	Offers a balanced perspective on model interpretability by emphasizing both local and global consistency of the trained model.
Monotonicity Metric [57, 27]	Emphasizes the importance of ensuring that the output behavior should be predictable and consistent with respect to specific input features.	Ensure that the model's predictions are reliable, especially in scenarios where erratic behavior can have significant consequences.
Pixel Flipping [51]	Investigates certain pixels or groups of pixels that have a pronounced influence on the classification outcome. Altering these pixels leads to changes in model predictions.	Ensures consistent and reliable interpretations at the pixel level, especially in domains where visual data is pivotal.
Region Perturbation [58]	Identifies critical regions in input data that have a disproportionate influence on the model's outcome. Offering a pathway to more interpretable models, especially in critical applications like medical imaging.	Ensures models focus on genuinely significant regions and do not base decisions on irrelevant features.

Table 4: The relevance of six properties in the context of robustness evaluation.

Property	**Key Contribution & Primary Assumptions**	**Influence & Impact**
Local Lipschitz Estimate [59]	Advocates for the robustness of explanations, arguing that similar inputs should produce similar explanations.	Highlights the importance of robust explanations in interpretability and introduces robustness metrics.
Max-Sensitivity [60]	Investigates objective evaluation measures of saliency explanations for DL models.	Sheds light on the fidelity and sensitivity aspects of saliency explanations.
Continuity [61]	Presents comprehensive methods for interpreting DNNs.	Focuses on the continuity and consistency of explanations as a measure of their robustness.
Consistency [62]	Proposes a framework to evaluate the faithfulness of local explanations for the underlying prediction model.	Highlights the consistency aspect of explanations as a measure of their robustness.
Relative Input Stability [63]	Addresses the stability of attribution-based explanation methods, crucial for model trustworthiness.	Rethinks the stability of input explanations, emphasizing their relevance in robustness evaluation.
Relative Representation Stability [63]	Focuses on the stability of attribution-based explanation methods to establish model trust.	Addresses stability of representational explanation for consistent interpretation.

Table 5: The relevance of six properties in the context of localization evaluation.

Property	**Key Contribution & Assumptions**	**Influence & Impact**
Pointing Game [64]	Top-down Neural Attention by Excitation Backprop in network hierarchies.	Proposes a method for generating task-specific attention maps in CNNs.
Attribution Localization [65]	Focusing on transparency in neural network's decisions using Layer-wise Relevance Propagation (LRP).	Contributes to the field of XAI by emphasizing transparency.
Top-K Intersection [66]	Interpretable semantic photo geolocation by estimating location in an image based on its content.	Delves into the nuances of image geolocation using image content.
Relevance Rank Accuracy [67]	Highlights the significance of ground truth evaluation of DL explanations with CLEVR-XAI.	Stress on the importance of ground truth evaluation for neural network explanations.
Focus [68]	Sheds light on potential biases in XAI methods.	Evaluates and rates XAI methods to uncover biases.
Receiver Operating Characteristics [69]	Introduction to ROC analysis are essential for visualizing and understanding classifier performance.	Addresses misconceptions and pitfalls in using ROC graphs.

Table 6: The relevance of three properties in the context of complexity evaluation.

Property	Key Contribution & Assumptions	Influence & Impact
Sparseness [73]	Highlights the significance of focusing on key features, minimizing the influence of irrelevant ones.	Uncover the influence of adversarial training on influencing model explanations and achieving sparseness.
Complexity [26]	Introduces quantitative evaluation metrics for feature-based model explanations: low sensitivity, high faithfulness, and low complexity.	Highlights the role of selection and aggregation of explanation functions in achieving desired complexity levels.
Effective Complexity [27]	Sets a benchmark for evaluating interpretability and provides a structured framework for understanding the nuances of effective complexity in explanations.	Advocates for objective metrics that encompass simplicity. Guides practitioners in discerning the balance between feature extraction and explanation.

Table 7: The relevance of two properties in the context of randomization evaluation.

Property	Key Contribution & Assumptions	Influence & Impact
Model Parameter Randomization [24]	Examines the role of adversarial training in influencing model explanations. A robust methodology for evaluating the effectiveness of various saliency methods, emphasizing sparseness and stability.	Highlights the importance of evaluating explanation techniques under randomized conditions and the potential misleading nature of visual assessments.
Random Logit Test [74]	Introduces the Cosine Similarity Convergence (CSC) metric to measure the information ignored from later layers, revealing that many explanations are independent of later layer parameters.	Addresses the potential discrepancies in explanations offered by various modified BP methods. Provides tools and metrics to evaluate the faithfulness of explanations under randomised scenarios.

the explanation in identifying the most influential regions or features in the input data. For example, in image-based tasks, high localization refers to the ability of a method to accurately pinpoint the exact regions in an image that contribute most significantly to the model's decision-making process [4].

In this context, Zhang et al. [64] pioneered the exploration of top-down attention mechanisms in CNNs, demonstrating how excitation backprop can be harnessed to generate task-specific attention maps. This is supported by the use of LRP to understand the nuances of a network's decisions [65]. In 2022, Theiner et al. [66] enter the intricate domain of image geolocation, recognizing the challenges and successes of determining locations solely from visual content. Their findings emphasize the crucial role CNNs play in semantic image interpretation. Arras et al. [67] emphasize the imperativeness of ground truth evaluation, rigorously assessing their accuracy beyond explanation generations.

An in-depth analysis of XAI methods reveals the inherent biases that contribute to the complexity of localization in the field. [68]. Fawcett et al.'s [69] exposition on the Receiver Operating Characteristics (ROC) analysis offers a holistic view of classifier performance visualization, marking a significant stride in understanding and evaluating model interpretability from a localization standpoint. The collective efforts outlined in Table 5 establish a nuanced narrative that emphasizes both challenges and advances in the interpretability of localization-centric models.

4. Complexity: The metric assesses the simplicity and comprehensibility of the generated explanation. Although detailed explanations can offer deeper insights, they may be overwhelming for certain end users. Striking a balance between detail and simplicity is pivotal.

In the evolving discourse on model interpretability, the realm of adversarial learning highlights the significance of two pivotal attributes:'sparseness' and 'stability' in explanations [73], asserting that a model's interpretation should zero in on the most crucial features, leaving out the redundant. Here sparseness implies minimizing attributions for irrelevant or weakly relevant features. Stability suggests that the explanations should remain consistent within a small local neighborhood of the input. Chalasani et al. [73] show that adversarial training with L_1-bounded adversaries results in models with sparse attribution vectors. Recently, [75] focused on investigating the interpretability of the Traffic Sign Recognition (TSR) model under adversarial attack conditions. Here, LIME and grad-CAM explanation proved more insightful in interpreting the model's behavior under adversarial conditions.

Parallelly, Bhatt et al. [26] present a pragmatic framework that focuses on quantitative metrics, with a unique emphasis on the intricacies of 'complexity' in feature-based model explanations. The authors introduce three primary criteria for evaluating explanations: low sensitivity, high faithfulness, and low complexity. They developed a framework for aggregating different explanation functions that prioritizes simplicity. The concept of 'effective complexity' also finds its roots in the work of Nguyen and Martínez [27]. Both emphasize the balance between the inherent subjectivity of interpretability and merits of objective metrics. Table 6 illustrates the multifaceted intricacies and imperatives of complexity in model interpretability.

5. Randomization: The metric assesses the reliability of an interpretable method by observing its behavior

Table 8: The relevance of three properties in the context of axiomatic evaluation.

Property	Key Contribution & Assumptions	Influence & Impact
Completeness [18]	Introduces two foundational axioms (Sensitivity and Implementation Invariance) and presents 'Integrated Gradients' as an attribution method	Provides a theoretical framework for evaluating attribution methods and critiques potential shortcomings of non-adherent methods
Non-Sensitivity [27]	Interpretability, while subjective, can be evaluated using objective measures like simplicity and broadness. Offers metrics to guide method selection, emphasizing distinct roles of feature extraction and explainability	Stresses the need for objective measures in interpretability, aligning with the axiomatic approach's foundational emphasis.
Input Invariance [76]	Examines potential pitfalls of saliency methods, proposing input invariance as a reliability benchmark. Reliable explanations shouldn't be sensitive to factors that do not influence the model prediction	Highlights the importance of genuine reflection of a model's decision-making in explanations

when the model's weights are randomized. A credible explanation method would produce non-informative or random explanations for a randomized model, indicating that the explanations are genuinely tied to the model's learned parameters [28].

The sanity check study [24] randomizes the model parameters to check how the explanation maps vary. Surprisingly, for some methods, such as guided backprop, the heatmaps remained unchanged. The study highlights the relevance of input features to neural network's prediction and relying solely on visual assessment can be misleading. Some saliency methods are found to be independent of both the model and the data-generating process. And, an analogy is drawn with edge detection in images, a technique that does not require training data or a model. An extended study on a wide range of modified backpropagation methods [74], including Deep Taylor Decomposition, LRP, Excitation Backpropagation, DeepLIFT [31], RectGrad [77], and Guided Backpropagation, showed a surprising result. Empirical evidence shows that the explanations provided by all of these methods, except DeepLIFT, remain independent of changes in the subsequent layers of the model. This phenomenon is attributed to the convergence of the relevance propagation to a rank-1 matrix. A rank-1 matrix essentially collapses the information to a single direction, making the backpropagated relevance vectors insensitive to the input. Consequently, even if the parameters in subsequent layers are randomized, the explanation generated by these techniques remain unaffected.

Unlike other methods, DeepLIFT [31] is not constrained by this limitation as it considers both positive and negative contributions by backpropagating the difference in activations relative to a reference activation. DeepLIFT preserves parameter dependency across the network, including the subsequent layers, resulting in more faithful explanations that reflect the overall network's workings. Collectively, the techniques outlined in Table 7 emphasize rigorous evaluation and discourage accepting explanations at face value.

6. Axiomatic Metrics: Axioms serve as foundational principles that any attribution method should ideally adhere to. This perspective is invaluable, because it offers strong theoretical groundwork for evaluating and comparing different explanation methodologies. The principles outlined in Table 8, including implementation invariance and sensitivity, offer a rigorous framework to test the validity and reliability of different interpretability techniques. Sundararajan et al. [18] critically assess existing attribution methods, emphasizing the absence of adherence to the introduced axioms.

In a related context, Nguyen and Martínez [27] address the challenges of quantitatively assessing interpretability. They suggested using objective metrics as an alternative to subjective evaluations, which can be unclear and open to interpretation. Regarding the dependability of saliency methods, Kindermans et al. [76] emphasize the concept of "input invariance." Together, these approaches underscore the multifaceted nuances of axiomatic evaluation, setting the stage for a comprehensive exploration of model transparency.

Incorporating these metrics into the evaluation process ensures a holistic assessment of interpretability methods, taking into account their theoretical foundations as well as their practical implications.

Conclusion

Numerous surveys have accompanied the recent surge in XAI research. In particular, gradient-based feature attribution methods, a cornerstone of XAI, have gained significant attention over the past decade. However, these surveys often offer broad overviews that lack extensive investigation of the specific techniques they deserve. This paper addresses this gap by providing a systematic review of the major advances in gradient-based explanations. We have meticulously detailed the "Gradient-based Visual Interpretation" section, providing crucial information to foster understanding and address key issues including the major applications in the summary tables.

Researchers usually evaluate derived explanations based on two aspects: explainability and faithfulness. Explain-

ability refers to the ability to make a model's decision understandable to humans. In particular, human assessment, fine-grain image recognition, and localization tests are adopted to determine whether the explanations align with user expectations. Faithfulness refers to the extent to which explanations accurately reflect internal decision-making processes. Specifically, various ablation tests are used to evaluate faithfulness from a causal perspective. In addition, randomization tests are employed to assess whether explanations are dependent on the model parameters and input instances. Other metrics such as algorithm efficiency and interactivity that are tailored to specific research challenges in XAI are not the focus of this paper and are therefore not included.

Also, due to the lack of a standard consensus on evaluating interpretability metrics, it is crucial to define and address specific evaluation challenges and methods. The most common being the hidden power of bias. Although the gradient term for input features may appear insignificant, model bias can play an outsized role in predictions. This can result in misleading explanations if we do not carefully factor in how bias propagates in the forward pass. We need to isolate the impact of bias to ensure that the explanations are truly representative of the input features. For example, with active ReLUs ($wx + b > 0$), bias plays a significant role in deeper layers alongside input features. However, feeding a zero-input image to a CNN activates very few ReLUs, minimizing the impact of bias on the final prediction. Understanding this relationship can be helpful for accurately interpreting the role of bias in model decisions.

Furthermore, the quality of gradient-based explanations can be highly dependent on the choice of hyperparameters. For example, the baseline point in IG [18] or the sampling steps in IDGI [41]. Finding the optimal settings is often counterintuitive and less guided by the input itself. It would be ideal to have explanation methods with fewer, easily understandable hyperparameters. In addition, many gradient-based techniques rely on assumptions about the connection between gradients and prediction logic. For example, we often attribute noisy visualizations to small gradients (the 'gradient saturation' problem), and this has led to techniques like IGs. However, the exact theoretical connection between noisy visualizations and smaller gradients remains unclear [52]. We still lack a strong explanation for why accumulating gradients helps identify important features. Although IG methods may seem to work well in practice, we need both theoretical and empirical evidence to fully justify their use and understand their limitations. A stronger theoretical foundation will ensure that these explanations are reliable under different conditions and facilitate the development of improved methods.

When discussing security risks, using gradients to explain a model's inner workings can unintentionally disclose excessive information, making the model susceptible to attacks [39]. Consider linear models, where gradients are essentially model weights. Sharing gradients as explanations is much like revealing the inner workings of the model. Research shows that extracting a model from its gradients is significantly more efficient than doing so from the prediction interface alone. It is critical to develop XAI techniques that balance explainability with model security to protect intellectual property.

Lastly, gradient-based explanations can be manipulated by adversarial attacks that subtly alter inputs without affecting the model's output. Attackers can manipulate input data with subtle changes imperceptible to humans that can significantly impact the explanation generated by the model. Say, a saliency map highlights specific image regions as crucial for a prediction. An attacker might modify those very regions slightly, without affecting the model's output, but causing the saliency map to pinpoint entirely different areas. The vulnerability of explanations may be attributed to high dimensionality and nonlinear nature of neural networks. A solution to this problem remains elusive.

Prioritizing the development of reliable explanation methods that can withstand adversarial attacks, along with countermeasures to audit such attacks, should be a top priority. Addressing these challenges will not only improve gradient-based methods, but also lay the foundation for fair, reliable, secure, and theoretically sound XAI techniques. This is particularly important when considering the implementation of these methods in sensitive applications.

Summary and Discussion

Every day, AI analyzes vast amounts of multimodal data—text, images, and even sounds, to make decisions. Our brains are amazing at interpreting the world around us, but a computer struggles with the same task? Like mistaking a turtle for a rifle [4]. This isn't just a quirky error; it challenges the reliability of decisions that affect everything, from our social media feeds to our car's navigation system. That is the power and the peril of AI.

Sometimes, decisions are hard, even for humans. Here, we see the unfair expectation for AI to be perfect, setting standards that no human could meet. We need to remember that AI reflects our world, including "uncertainties". In the quest for clarity, researchers worldwide have been developing methods to make AI's decisions transparent, scrambling to open the AI black-box with no clear consensus. But is a simpler story the truth? We need to know which XAI technique is accountable and which explanation is truly robust.

As AI becomes more intertwined with our daily activities and neural networks become increasingly complex, tools and techniques for interpretability must keep pace. Some of the recent advances in the field include hybrid methods that combine the strengths of multiple XAI tech-

niques, domain-adaptive interpretability methods tailored for specific industries like finance or healthcare, and meta-interpretability frameworks that offer insights into the behavior of XAI methods themselves. Through this comprehensive survey, we have navigated the landscape of XAI, shedding light on propagation-based methods and their significance in the broader context of interpretability. Figure 1 presents the first comprehensive timeline summarizing the major publication history of gradient-based explanation methods. This visualization, along with the summary provided in Table 1 enables researchers to comprehend how each new algorithm builds upon the strengths and weaknesses of its predecessors, fostering a clear understanding of the field's chronological development. By tracing this trajectory of critical thinking, researchers can identify the latest advances and emerging trends, ultimately inspiring further contributions to the field of gradient-based explanations. The field continues to evolve and grow as new paradigms of explainability emerge in the future.

Conflict of interest

Authors state no conflict of interest.

Acknowledgments

This work was supported by the Research Council of Norway Project (nanoAI, Project ID: 325741), H2020 Project (OrganVision, Project ID: 964800), HORIZON-ERC-POC Project (Spermotile, Project ID: 101123485), and VirtualStain (UiT, Cristin Project ID: 2061348).

References

1. O'Sullivan S, Janssen M, Holzinger A, Nevejans N, Eminaga O, Meyer C, and Miernik A. Explainable artificial intelligence (XAI): closing the gap between image analysis and navigation in complex invasive diagnostic procedures. World Journal of Urology 2022 :1–10
2. Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K, and Ghayvat H. CNN variants for computer vision: History, architecture, application, challenges and future scope. Electronics 2021; 10:2470
3. Adadi A and Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access 2018; 6:52138–60
4. Somani A, Horsch A, and Prasad DK. Interpretability in Deep Learning. Springer International Publishing, 2023. DOI: 10.1007/978-3-031-20639-9
5. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, and Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. Nature communications 2019; 10:1–8
6. Lipton ZC. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 2018; 16:31–57
7. Regulation GDP. General data protection regulation (GDPR). Intersoft Consulting 2018; 24
8. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019; 1:206–15
9. Dwivedi R, Dave D, Naik H, Singhal S, Omer R, Patel P, Qian B, Wen Z, Shah T, Morgan G, et al. Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys 2023; 55:1–33
10. Minh D, Wang HX, Li YF, and Nguyen TN. Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review 2022 :1–66
11. Simonyan K, Vedaldi A, and Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 2013
12. Smilkov D, Thorat N, Kim B, Viégas F, and Wattenberg M. Smoothgrad: removing noise by adding noise. arXiv preprint arXiv:1706.03825 2017
13. Ribeiro MT, Singh S, and Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016 :1135–44
14. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, and Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE international conference on computer vision*. 2017 :618–26
15. Leino K, Sen S, Datta A, Fredrikson M, and Li L. Influence-directed explanations for deep convolutional networks. *2018 IEEE international test conference (ITC)*. IEEE. 2018 :1–8

16. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems 2017; 30

17. Springenberg JT, Dosovitskiy A, Brox T, and Riedmiller M. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806 2014

18. Sundararajan M, Taly A, and Yan Q. Axiomatic attribution for deep networks. *International conference on machine learning*. PMLR. 2017 :3319–28

19. Baehrens D, Schroeter T, Harmeling S, Kawanabe M, Hansen K, and Müller KR. How to explain individual classification decisions. The Journal of Machine Learning Research 2010; 11:1803–31

20. Supekar K, Los Angeles C de, Ryali S, Cao K, Ma T, and Menon V. Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism. The British Journal of Psychiatry 2022; 220:202–9

21. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, and Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nature Reviews Genetics 2023; 24:125–37

22. Rebuffi SA, Fong R, Ji X, and Vedaldi A. There and back again: Revisiting backpropagation saliency methods. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020 :8839–48

23. Amorim JP, Abreu PH, Santos J, Cortes M, and Vila V. Evaluating the faithfulness of saliency maps in explaining deep learning models using realistic perturbations. Information Processing & Management 2023; 60:103225

24. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, and Kim B. Sanity checks for saliency maps. Advances in neural information processing systems 2018; 31

25. Binder A, Weber L, Lapuschkin S, Montavon G, Müller KR, and Samek W. Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023 :16143–52

26. Bhatt U, Weller A, and Moura JM. Evaluating and aggregating feature-based model explanations. *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021 :3016–22

27. Nguyen Ap and Martínez MR. On quantitative aspects of model interpretability. arXiv preprint arXiv:2007.07584 2020

28. Hedström A, Weber L, Krakowczyk D, Bareeva D, Motzkus F, Samek W, Lapuschkin S, and Höhne MMC. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. Journal of Machine Learning Research 2023; 24:1–11

29. Agarwal C, Krishna S, Saxena E, Pawelczyk M, Johnson N, Puri I, Zitnik M, and Lakkaraju H. Openxai: Towards a transparent evaluation of model explanations. Advances in Neural Information Processing Systems 2022; 35:15784–99

30. Shrikumar A, Greenside P, Shcherbina A, and Kundaje A. Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 2016

31. Shrikumar A, Greenside P, and Kundaje A. Learning important features through propagating activation differences. *International conference on machine learning*. PMLR. 2017 :3145–53

32. Adebayo J, Gilmer J, Goodfellow I, and Kim B. Local explanation methods for deep neural networks lack sensitivity to parameter values. arXiv preprint arXiv:1810.03307 2018

33. Srinivas S and Fleuret F. Full-gradient representation for neural network visualization. Advances in neural information processing systems 2019; 32

34. Erion G, Janizek JD, Sturmfels P, Lundberg SM, and Lee SI. Learning explainable models using attribution priors. 2019

35. Xu S, Venugopalan S, and Sundararajan M. Attribution in scale and space. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020 :9680–9

36. Janizek JD, Sturmfels P, and Lee SI. Explaining explanations: Axiomatic feature interactions for deep networks. Journal of Machine Learning Research 2021; 22:1–54

37. Kapishnikov A, Venugopalan S, Avci B, Wedin B, Terry M, and Bolukbasi T. Guided integrated gradients: An adaptive path method for removing noise. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021 :5050–8

38. Pan D, Li X, and Zhu D. Explaining deep neural network models with adversarial gradient integration. *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*. 2021

39. Wang Z, Fredrikson M, and Datta A. Robust Models Are More Interpretable Because Attributions Look Normal. *International Conference on Machine Learning*. PMLR. 2022 :22625–51

40. Bykov K, Hedström A, Nakajima S, and Höhne MMC. Noisegrad—enhancing explanations by introducing stochasticity to model weights. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 6. 2022 :6132–40

41. Yang R, Wang B, and Bilgic M. IDGI: A framework to eliminate explanation noise from integrated gradients. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023 :23725–34

42. Wang H, Wang Z, Du M, Yang F, Zhang Z, Ding S, Mardziel P, and Hu X. Score-CAM: Score-weighted visual explanations for convolutional neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020 :24–5

43. Zhang Q, Wu YN, and Zhu SC. Interpretable convolutional neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018 :8827–36

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

45. Tjoa E and Guan C. A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE transactions on neural networks and learning systems 2020; 32:4793–813

46. Zeiler MD, Taylor GW, and Fergus R. Adaptive deconvolutional networks for mid and high level feature learning. *2011 international conference on computer vision*. IEEE. 2011 :2018–25

47. Chattopadhay A, Sarkar A, Howlader P, and Balasubramanian VN. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. 2018 :839–47

48. García-Torres J and Meinich-Bache Ø. Interpretability in Video-based Human Action Recognition: Saliency Maps and GradCAM in 3D Convolutional Neural Networks. 2024

49. Hesse R, Schaub-Meyer S, and Roth S. Fast axiomatic attribution for neural networks. Advances in Neural Information Processing Systems 2021; 34:19513–24

50. López A, Zobolas J, Lingjærde OC, Nebdal D, Fleischer T, and Aittokallio T. Explainable multi-omics deep clustering model reveals an important role of DNA methylation in pancreatic ductal adenocarcinoma. 2024

51. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, and Samek W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one 2015; 10:e0130140

52. Wang Y, Zhang T, Guo X, and Shen Z. Gradient based Feature Attribution in Explainable AI: A Technical Review. arXiv preprint arXiv:2403.10415 2024

53. Seo J, Choe J, Koo J, Jeon S, Kim B, and Jeon T. Noise-adding methods of saliency map as series of higher order partial derivative. arXiv preprint arXiv:1806.03000 2018

54. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, and Lee SI. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence 2020; 2:56–67

55. Waa J van der, Nieuwburg E, Cremers A, and Neerincx M. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence 2021; 291:103404

56. Alvarez Melis D and Jaakkola T. Towards robust interpretability with self-explaining neural networks. Advances in neural information processing systems 2018; 31

57. Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, Houde S, Liao QV, Luss R, Mojsilović A, et al. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012 2019

58. Samek W, Binder A, Montavon G, Lapuschkin S, and Müller KR. Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems 2016; 28:2660–73

59. Alvarez-Melis D and Jaakkola TS. On the robustness of interpretability methods. arXiv preprint arXiv:1806.08049 2018

60. Yeh CK, Hsieh CY, Suggala A, Inouye DI, and Ravikumar PK. On the (in) fidelity and sensitivity of explanations. Advances in Neural Information Processing Systems 2019; 32

61. Montavon G, Samek W, and Müller KR. Methods for interpreting and understanding deep neural networks. Digital signal processing 2018; 73:1–15

62. Dasgupta S, Frost N, and Moshkovitz M. Framework for evaluating faithfulness of local explanations. *International Conference on Machine Learning*. PMLR. 2022 :4794–815

63. Agarwal C, Johnson N, Pawelczyk M, Krishna S, Saxena E, Zitnik M, and Lakkaraju H. Rethinking stability for attribution-based explanations. *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*. 2022

64. Zhang J, Bargal SA, Lin Z, Brandt J, Shen X, and Sclaroff S. Top-down neural attention by excitation backprop. International Journal of Computer Vision 2018; 126:1084–102

65. Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W, and Lapuschkin S. Towards best practice in explaining neural network decisions with LRP. *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020 :1–7

66. Theiner J, Müller-Budack E, and Ewerth R. Interpretable semantic photo geolocation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022 :750–60

67. Arras L, Osman A, and Samek W. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. Information Fusion 2022; 81:14–40

68. Arias-Duart A, Parés F, Garcia-Gasulla D, and Giménez-Ábalos V. Focus! rating XAI methods and finding biases. *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2022 :1–8

69. Fawcett T. An introduction to ROC analysis. Pattern recognition letters 2006; 27:861–74

70. Ancona M, Ceolini E, Öztireli C, and Gross M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. *International Conference on Learning Representations*. 2018. Available from: https://openreview.net/forum?id=Sy21R9JAW

71. Rieger L and Hansen LK. IROF: a low resource evaluation metric for explanation methods. *Workshop AI for Affordable Healthcare at ICLR 2020*. 2020

72. Reyes M, Meier R, Pereira S, Silva CA, Dahlweid FM, Tengg-Kobligk Hv, Summers RM, and Wiest R. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology: artificial intelligence 2020; 2:e190043

73. Chalasani P, Chen J, Chowdhury AR, Wu X, and Jha S. Concise explanations of neural networks using adversarial training. *International Conference on Machine Learning*. PMLR. 2020 :1383–91

74. Sixt L, Granz M, and Landgraf T. When explanations lie: Why many modified bp attributions fail. *International Conference on Machine Learning*. PMLR. 2020 :9046–57

75. Khan MA and Park H. Exploring Explainable Artificial Intelligence Techniques for Interpretable Neural Networks in Traffic Sign Recognition Systems. Electronics 2024; 13:306

76. Kindermans PJ, Hooker S, Adebayo J, Alber M, Schütt KT, Dähne S, Erhan D, and Kim B. The (un) reliability of saliency methods. Explainable AI: Interpreting, explaining and visualizing deep learning 2019 :267–80

77. Kim B, Seo J, Jeon S, Koo J, Choe J, and Jeon T. Why are saliency maps noisy? cause of and solution to noisy saliency maps. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019 :4149–57