2023 IEEE 27th International Conference on Intelligent Engineering Systems (INES) | 979-8-3503-2851-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/INES59282.2023.10297629

Evaluation Metrics for XAI: A Review, Taxonomy, and Practical Applications

Md Abdul Kadir
German Research Center for Artificial Intelligence
Saarbrücken, Germany
abdul.kadir@dfki.de

Amir Mosavi
Obuda University
Budapest, Hungary
amir.mosavi@uni-obuda.hu

Daniel Sonntag
German Research Center for Artificial Intelligence
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

***Abstract*—Within the past few years, the accuracy of deep learning and machine learning models has been improving significantly while less attention has been paid to their responsibility, explainability, and interpretability. eXplainable Artificial Intelligence (XAI) methods, guidelines, concepts, and strategies offer the possibility of models' evaluation for improving fidelity, faithfulness, and overall explainability. Due to the diversity of data and learning methodologies, there needs to be a clear definition for the validity, reliability, and evaluation metrics of explainability. This article reviews evaluation metrics used for XAI through the PRISMA systematic guideline for a comprehensive and systematic literature review. Based on the results, this study suggests two taxonomy for the evaluation metrics. One taxonomy is based on the applications, and one is based on the evaluation metrics.**

***Keywords*—XAI, machine learning, deep learning, explainable artificial intelligence, explainable AI, explainable machine learning; metrics; evaluation**

I. Introduction

Explainable Artificial Intelligence (XAI[1]) focuses on understanding the reasoning behind machine learning and deep learning models' decisions across a range of AI applications[2, 3]. XAI's goal is to aid users in building a comprehensive and accurate understanding of these algorithms, fostering confidence in their outputs[4, 5, 6]. Despite the many methods for explainability, researchers still lack consensus on the precise nature and practical properties of an Explanation[7]. Future research should define explainability and develop structured formats of Explanations, accommodating as many aspects as possible. Explainability in psychology, tied to trust, transparency, and privacy, identifies humans as final explanation recipients[8]. This emphasizes the necessity of interactive visual explanations in XAI and the need for psychometric research. Various studies in AI sub-domains call for fundamental research on explainability measurement[9, 10, 11].

Research has focused on explaining methods with neural and Bayesian networks and extracting rule clusters, aiming to generate human-interpretable rules without sacrificing accuracy [12]. Despite several attempts to survey and categorize explainability methods [13], there's consensus on the need for explanations to be comprehensible to laypeople and provide actionable information [14, 15, 16]. These studies underscore the need for systematic analysis of explainability metrics. This paper aims to systematically review XAI studies, focusing on articles that conceptually and theoretically address explainability and propose methods for evaluating XAI. The framework found that explainers develop explainability methods evaluated using metrics, often based on complex models or numerical approximations [17, 18, 19, 20]. However, the absence of robust evaluation metrics may underestimate the correlation between a trained model and its visual explainer, leading to potential inaccuracies in explanations. Hence, it is crucial to develop a reliable evaluation metric for ensuring high-quality, compelling, and informative visual explanations.

II. Methodology

The methodology, following the PRISMA guideline, integrates a comprehensive literature search with systematic screening. The primary database is Scopus[1], supplemented by Google Scholar[2] for additional or missing literature. Arxiv[3] is also utilized for early-version articles in AI. Initial queries, including 'explainable artificial intelligence', 'XAI', 'explainable AI', 'explainable machine learning', and 'explainable deep learning', resulted in 6122 articles on various explainable ML and deep learning methods and applications. Following the PRISMA[4] guideline, the research methodology, depicted in Fig. 1, refines the initial 6122 documents down to pertinent articles on XAI evaluation metrics. The challenge lies in the

This research is supported by DFKI IML (iml.dfki.de) and the Ministry of Science and Culture of the German State of Lower Saxony. For more information, visit cst.dfki.de. DFKI is supported by the Lower Saxony Ministry of Science and Culture. The responsibility lies with the authors.

[1] https://www.scopus.com/home.uri
[2] https://scholar.google.com
[3] https://arxiv.org
[4] http://www.prisma-statement.org/

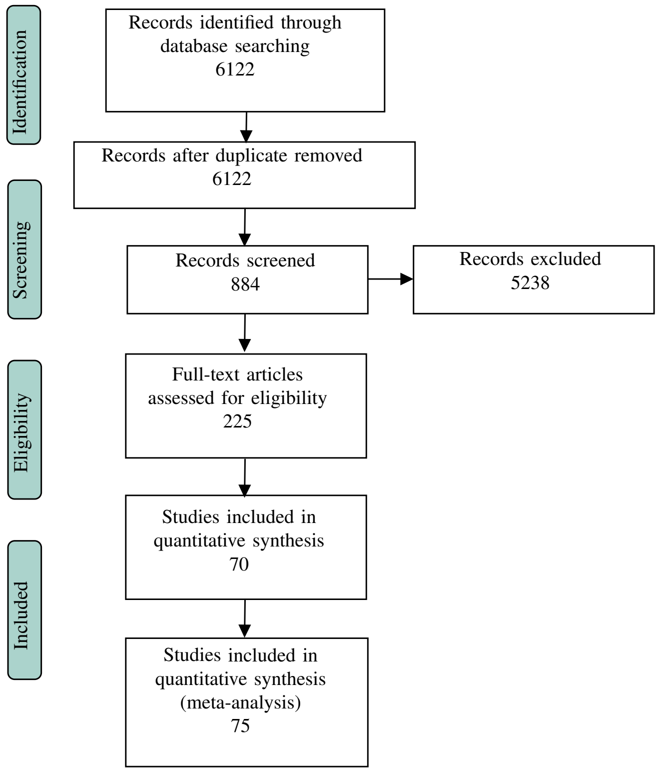

Fig. 1. The research methodology flow diagram follows the PRISMA guidelines to identify relevant literature and screen it to narrow down the amount of literature.

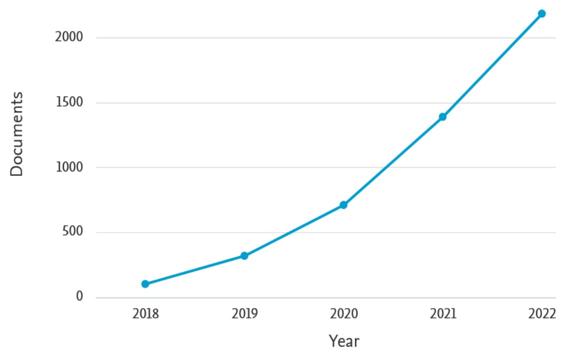

Fig. 2. The initial queries for XAI literature resulted in 6122 articles. The graph also indicates an upward trend in XAI research.

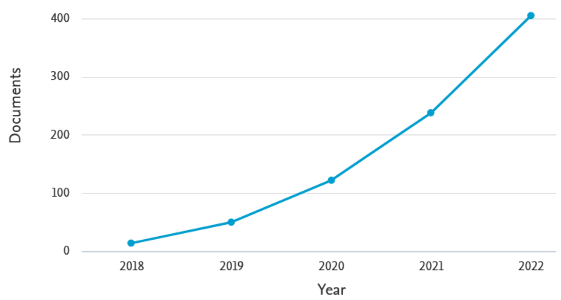

Fig. 3. After the first screening for explainability evaluation metrics, the literature review revealed that there were less than half a thousand research articles containing explanation metrics in 2022. However, there is an upward trend in the use of explanation metrics in research.

diverse definitions of explainability across applications and domains, and there's no common keyword for efficient screening. Therefore, we utilized a broad list of keywords and phrases typically associated with explainability. It's noteworthy that the terms 'evaluation metrics' or 'metrics' alone are insufficient for identifying relevant articles, due to diverse communication of explainability. Thus, combining these with explainability-related keywords effectively filters out irrelevant articles.

Consequently, screening the keywords of metrics, evaluation, and explainability through the entire fields of the articles results in 884 documents. Following keywords related to explainability were explored in this stage, i.e., *Actionability*, *Transparency*, *Transferability*, *Completeness*, *Satisfaction*, *Sensitivity*, *Stability*, *Informativeness*, *Robustness*, *Understandability*, *Monotonicity*, *Comprehensibility*, *Correctability*, *Interpretability*, *Efficiency*, *Explicability*, *Explicitness*, *Faithfulness*, *Intelligibility*, *Interactivity*, *Interestingness* after further screening the titles, abstract and keywords the number of relevant articles is reduced to 225. A secondary screening, involving in-depth reading, eliminated 155 articles focusing on metrics unrelated to explainability, often measuring performance and briefly discussing explainability. This screening further reduced the relevant articles to 75. Subsequent selection of original and highly relevant studies led to the final 70 articles, which are categorized according to the metrics presented in the tables.

III. Result

XAI research has seen rapid expansion over the past five years, as depicted in Fig. 2 based on the number of published articles. The initial search yielded 6122 XAI-focused articles. However, only a small portion included evaluation metrics for measuring explainability—critical for assessing explanation quality—amounting to about 884 articles. The distribution of these results over the past five years is shown in Fig. 3.

Table I lists key studies on Explainable Artificial Intelligence (XAI) from 2021 to 2023. These papers explore various XAI themes, tools, and applications across sectors like healthcare, education, and industry. Techniques utilized include deep learning, knowledge distillation, and statistical testing to build accessible AI models. Various assessment metrics, algorithms, and XAI tools such as SHAP, LIME, and LEAF were used to enhance model interpretability. Overall, the studies aim to enhance human-AI collaboration and address AI implementation challenges. As AI applications increase, so does the need for explainability, leading to proposed evaluation metrics for AI-generated explanations.

Chinu and Bansal [21] highlights the explanation metric's importance in assessing relief application responses. Schwalbe and Finzel [13] notes the growing popularity of explainable techniques and metrics. In power quality distribution, explain-

TABLE I
THIS TABLE PROVIDES A SUMMARY OF THE MOST RELEVANT LITERATURE THAT APPLIED XAI TECHNIQUES IN SOLVING PROBLEMS WITH REAL-WORLD DATA. IT HIGHLIGHTS THE RESEARCH THAT HAS UTILIZED XAI TECHNIQUES AND THEIR PRACTICAL APPLICATIONS.

Reference	Method	Application
[21] 2023	Explainable AI: To Reveal the Logic of Black-Box Models	Interpretable; Transparency; Quality metrics;
[22] 2023	A taxonomy for XAI methods	XAI; Interpretability; Meta-analysis
[23] 2023	XAI Evaluation metric: Traceability rate	Drug recommendation; Explainability; Traceability
[24] 2023	Explaining Machine Learning Model Explanations	Interpretability; GUI for Explanation
[25] 2022	XAI methods evaluation metric	Educational data; Learning analytics
[26] 2022	Multi-modal image-fusion model knowledge distillation and explainable AI	XAI in Medicine; Image generation
[27] 2022	Measuring Explainability and Trustworthiness of Power Quality Disturbances Classifiers	XAI in Power; Power quality disturbances (PQDs)
[28] 2022	A New Explainable Deep Learning Framework for Cyber Threat Discovery	Anomaly detection; IIoT; industrial networks
[29]2022	Putting explainable AI in context: institutional explanations for medical AI	AI and health; Epistemic risk; Ethical design
[30] 2022	Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance	Predictive maintenance
[31] 2022	XAI in IC defect detection	Explainable arch.; Hierarchical clustering
[32] 2020	Knowledge-Aware eXplainable AI	Knowledge-base;
[33] 2022	A human-agent architecture for explanation formulation	HCI; Multi-agent systems
[12] 2021	Notions of explainability and evaluation approaches for explainable artificial intelligence	Evaluation methods; Notions of explainability
[34] 202	Explainable artificial intelligence for bias detection	Computerised Tomography
[35] 2021	LEAF to evaluate local linear XAI methods	Local explanation; ML Auditing
[36] 2021	XAI in anomaly detection	Anomaly detection; Cryptomining
[37] 2021	XAI for Default Privacy Setting Prediction	Privacy preference
[38] 2022	Explaining AI with Narratives	Explainability, NL
[39] 2021	Interaction with Explanations	User interaction
[38] 2022	A survey on improving NLP models with human explanations	User interaction
[40] 2020	Explanatory Interactive Image Captioning	Image captioning

ability metrics help ensure reliable decisions, says Machlev et al. [27]. [28] underlines the explanation metric's role in detecting IoT data anomalies for security enhancement. Despite extensive research, a consensus on explanation definition and assessment is needed, according to Vilone and Longo [41]. While many works contribute to this field, Li et al. [32] notes the need for a clear taxonomy and systematic review. Further, Mualla et al. [33], Li et al. [42] propose new explanation techniques, reusing LIME's metric for evaluation.

Meanwhile, Palatnik de Sousa et al. [34] argue that performance metrics achieved by AI models can give users the impression that there is no bias. Hence, explaining classification and evaluating the explanation based on proper metrics is necessary. Additionally, Amparore et al. [35] addresses the problem of identifying a clear and unambiguous set of metrics for evaluating Local Linear Explanations. They also propose a LEAF framework for explanation evaluation to end-users.

Finally, for practical medical applications, Theunissen and Browning [29] suggests that metrics for evaluating post-hoc explanations are necessary. The metrics should evaluate the accuracy of the explanation, and there should be procedures for auditing the system to prevent biases and failures from going unaddressed. In summary, various researchers have proposed different metrics and frameworks to evaluate the quality of explanations produced by AI models. While there is still no consensus on how to define and evaluate explanations and explainability metrics' importance in understanding AI models and ensuring trustworthy decision-making cannot be overstated.

In our recent review of application-related research, we have identified that the evaluation technique is not the sole focus of interest but rather the explanation method itself. We found that in many cases, explanation evaluation was only qualitatively assessed, and the quality of the explanation was taken for granted without using any specific evaluation technique. However, several terminologies were reintroduced, such as local explanation, attribute, post-hoc explanation, sensitivity, trustworthiness, causal interpretation, traceability, and auditing. We discovered that sensitivity measurement was used frequently in the literature. This method is closely related to the taxonomy in 6f Fig. III. The sensitivity measurement evaluates the impact of input features on the model's output, which helps to identify the most critical features. It allows us to understand the contribution of each input feature to the model's prediction and to evaluate the explanation's quality. However, other terminologies, such as trustworthiness, causal interpretation, traceability, and auditing, can provide additional insights into the explanation's reliability and usability.

An alternative taxonomy is proposed in FIg. 5. In our recent review of application-related research, we have identified that the evaluation technique is not the sole focus of interest but rather the explanation method itself. We found that in many cases, explanation evaluation was only qualitatively assessed, and the quality of the explanation was taken for granted without using any specific evaluation technique. However, several terminologies were reintroduced, such as local expla-

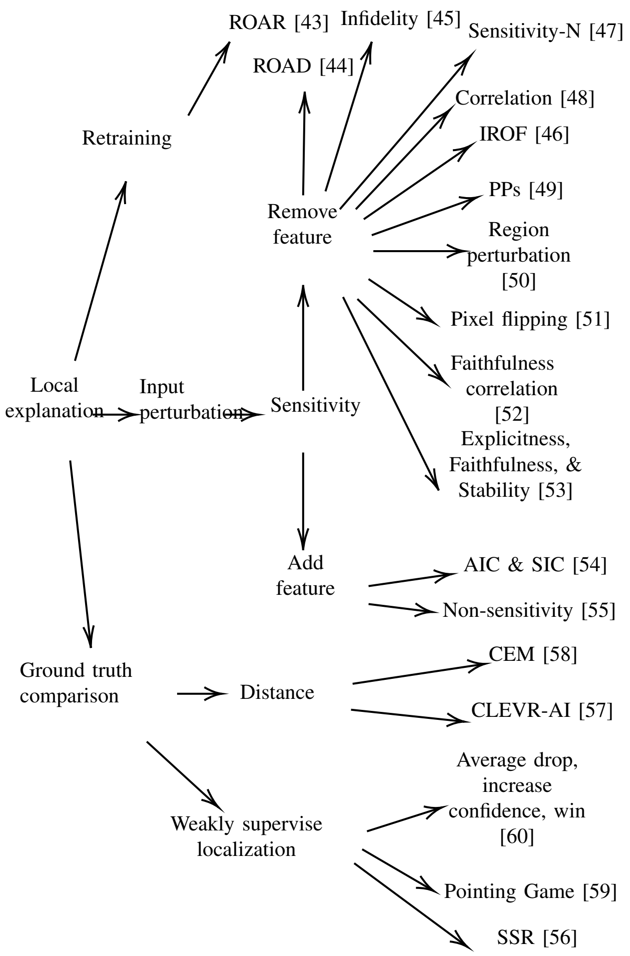

Fig. 4. Proposed taxonomy based the methodologies of the explainability evaluation

nation, attribute, post-hoc explanation, sensitivity, trustworthiness, causal interpretation, traceability, and auditing. We discovered that sensitivity measurement was used frequently in the literature. This method is closely related to the taxonomy in Fig. 5. The sensitivity measurement evaluates the impact of input features on the model's output, which helps to identify the most critical features. It allows us to understand the contribution of each input feature to the model's prediction and to evaluate the explanation's quality. However, other terminologies, such as trustworthiness, causal interpretation, traceability, and auditing, can provide additional insights into the explanation's reliability and usability.

In Table I, we found that local explanation is necessary for plenty of applications. A local explanation can be defined as an explanation that we get individual basis based on each decision the model makes. They can be post-hoc and generated after deploying a machine-learning model. Evaluation of local explanation can be done in three ways. After removing the relevant feature from the dataset based on the explanation, they retrained a proxy model after evaluating the new model's performance on untouched test data. Suppose the test accuracy

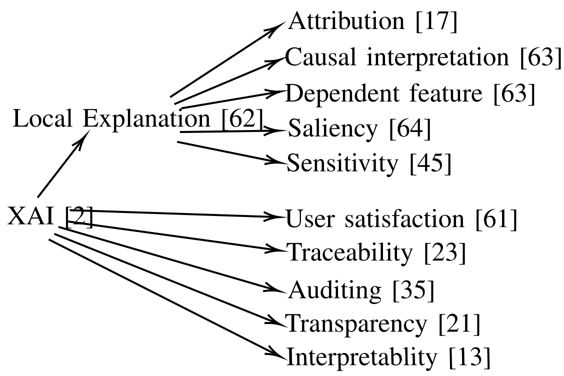

Fig. 5. Proposed taxonomy based on the XAI applications

of the newly created model is lower than the original model's accuracy. In that case, training on data with missing features creates an entirely different model than the original model. It signifies that the features removed from training data contribute to the original model's decision. This method has high computational demand due to the retraining process.

The second approach is ground-truth-based evaluation, and the explanation is compared with the ground-truth explanation data. Different distance metrics are used to identify how far the explanation is from the ground truth. Ground truth can also be user feedback on a model's local explanation [57, 65]. Some researchers used weakly supervised localization techniques to see how a saliency-based explanation can be meaningful to localize an object in an image [56, 60]. They proposed some metrics called SSR, Point Game, Average drop, Increase confidence, and Win. Kapishnikov et al. [54], Rguibi et al. [66] used Accuracy Information Curves (AICs), Softmax Information Curves (SICs), and Performance Information Curves PICs XAI evaluation. [67, 68, 69] has used the area under perturbation curve (AOPC) for understanding the decisions of CNN using MoRF curve and evaluates the explainability of their proposed model. Recently, Veldhuis et al. [70] leveraged explainable AI methods for DNA analysis. Xi et al. [23], Apicella et al. [71, 72], Schinle et al. [73] reported their experiment results with MoRF curve or its variations as reliable evaluation metrics.

There has been tremendous interest in unsupervised techniques for evaluating explanations in the last decade. Most of these methods work based on removing or adding information from the input data and measuring the changes in the output of the mode. SIC and AIC scores [54] Non-sensitivity [55] scores are measured based on the output of the model. When data is fed to an input, the output scores represent the influence of essential and nonimportant features in the model output. Similarly, removing features from input data also influences the model. Sensitivity-N can measure the influence [74], and Faithfulness Correlation [52]. The feature removal from the input is a tricky process, and the feature removal should have the property of missingness [75]. Such algorithms are also proposed by [46, 67].

A. Sensitivity analysis

Explanation sensitivity refers to how much a machine learning model's output is affected by different types of explanations or interpretability methods applied to it. In other words, it measures how much the output of a model changes when different explanations are provided for it. Sensitivity analysis is a key part of explainable AI and helps researchers and practitioners understand how reliable and robust the explanations of machine learning models are. Table II represents the sensitivity analysis methods used for the evaluation of the XAI methods. The definition of classic explanation sensitivity [45] can be expressed as follows: For any $j \in \{1, ..., d\}$,

$$[\nabla_x \phi(f(x))]_j = \lim_{\epsilon \to 0} \frac{\phi(f(x + \epsilon e_j)) - \phi(f(x))}{\epsilon} \quad (1)$$

where $e_j \in R_d$ is the j^{th} coordinate basis vector, with j^{th} entry one and all others zero. It quantifies how the explanation changes as the input is varied infinitesimally where f is the model, ϕ is the explainer e_j is the changes in the input features and ϵ is the deviation.

Table II includes various research articles that employ Explainable Artificial Intelligence (XAI) methods in different applications. Sensitivity Analysis is one of the XAI methods used to analyze the impact of input features on the model's output. Some of the applications include Covid-19 diagnosis, self-driving cars, brain-computer interface systems, seismic facies classification, predicting the functional impact of gene variants, discovering bias in structured pattern classification datasets, smart agriculture, compression, feature selection, volcano detection, optical water types, feature importance analysis, threat detection, survival analysis, and COVID-19 screening using chest X-ray images. The XAI methods employed in these applications include LIME, SHAP, Multi-Objective Sensitivity Pruning, Graph embedding, Grad-CAM, Gaussian processes, Hierarchical Interpretable models, Attack trees, Bayesian networks, and Grad-CAM++. They employed an explanation evaluation technique for evaluating the output of the explanation methods. For example, Kim and Joe [77] used sensitivity analysis for evaluating explanations in self-driving cars' decision-making process. In anomaly detection explanation, sensitivity can be used to evaluate the model's decision [93].

B. Faithfulness Correlation and Faithfulness Estimate metrics

Faithfulness correlation measures the linear relationship between the model predictions and the training data. It quantifies how well the model can capture the patterns and relationships in the training data. A high faithfulness correlation indicates that the model faithfully detects patterns and information in the training data. In contrast, a low faithfulness correlation indicates that the model may be over-fitting or under-fitting the data. Faithfulness Correlation [52] iteratively replaces a random subset of given attributions with a baseline value. Then it measures the correlation between the attribution subset and the difference in function output. On the other hand, Faithfulness Estimate [53] computes the correlation between

TABLE II
SENSITIVITY ANALYSIS FOR XAI METHODS

Reference	Method	Application
[76] 2022	Covid-MANet	Sensitivity analysis; Lesion localisation
[77] 2022	An XAI method for convolutional neural networks	Self driving car; CNN; Sensitivity of features
[78] 2022	XAI in brain-computer interface systems	Brain–computer
[79] 2022	Quantifying the sensitivity of seismic facies classification to seismic attribute selection	Sensitivity of seismic attributes; Seismic geomorphology
[80] 2022	Predicting KCNQ1 variants with ANN	Protein structure
[81] 2022	Discover bias	Understanding bias; Fairness
[82] 2022	Explainable AI at Work! What Can It Do for Smart Agriculture?	Explainability in Agriculture data
[83] 2022	MOSP: Pruning of Deep Neural Networks	Neural network compression
[84] 2022	A Feature Selection Method via Graph Embedding and Global Sensitivity Analysis	Feature engineering
[85] 2022	XAI in Detection of VDS	Volcanic Deformation analysis
[86] 2022	Learning Relevant Features of Optical Water Types	See water
[87] 2021	Deep belief network framework and its application for feature importance analysis	Feature engineering
[72] 2021	Explanations in terms of Hierarchically organised Middle Level Features	Feature understanding
[88] 2021	Adversarial policy training against deep reinforcement learning	Preventing adversarial attacks
[89] 2021	Efficient Estimation of the ANOVA Mean Dimension, with an Application to Neural Net Classification	Dimensionality reduction
[90] 2021	Bayesian Networks for Online Cybersecurity Threat Detection	Threat detection and analysis
[91] 2020	Explaining unreliable ML survival models	Reducing data demand
[92] 2020	Evaluation of scalability and degree of fine-tuning	Medical imaging; Low training data
[93] 2020	A deep Taylor decomposition of one-class models	Outlier detection; Unsupervised learning
[63] 2020	Interpretable ML A Brief History	Dependent features; Causal interpretation
[94] 2019	XAI NLP; Biomedical Classification	Drawback of blackbox model

probability drops and attribution scores on various points. Table III summarizes XAI studies, including the Faithfulness Correlation and Faithfulness Estimate metrics. According to [52], the faithfulness of an explanation function g to a predictor f at a point x with a subset size of $|S|$ is defined as follows:

$$\mu_F(f, g; x) = \underset{S \in \binom{[d]}{|S|}}{corr} \left(\sum_{i \in S} \left(g(f, x)_i, f(x) - f(x_{[x_s = \hat{x}_s]}) \right) \right) \tag{2}$$

d is the dimension of x. x_s are particular features to a baseline value $\hat{x}_s$. Table III lists various studies and research papers that showcase the application of faithfulness metrics in explainable AI (XAI). Faithfulness is one of the essential metrics used to evaluate the performance of XAI methods. It measures how well an AI model's explanations align with its underlying decision-making processes. For instance, in the medical image analysis study by Jin et al. [14], the authors proposed guidelines to evaluate the faithfulness of clinical XAI models. Similarly, the G-LIME method introduced by Li et al. [42] aims to provide interpretable deep learning by ensuring the faithfulness of local interpretations of deep neural networks using global priors. Other studies in the table that utilize faithfulness metrics include those in autonomous driving and natural language processing. These studies illustrate the significance of faithfulness in XAI and its application across different domains.

C. Monotonicity Metric

Monotonicity Metric introduced by Luss et al. [49] generates contrastive explanations with monotonic attribute functions. Arya et al. [48] further elaborates on these metrics. It starts from a reference baseline to incrementally place each feature on the baseline surface from a sorted attribution vector, measuring the effect on model performance. Recently Monotonicity Metric has been employed by several studies [112, 113, 114, 115].

D. Pixel Flipping

Pixel Flipping [51] captures the impact of perturbing pixels in descending order according to the attributed value on the classification score. Wullenweber et al. [116], Pitroda et al. [117] used Pixel Flipping metric for evaluating explanations for the predictions of COVID-19 cough classifiers and lung disease classification.

$$d_k(p) = \frac{\sum_{N \in digits(k)} N(p)}{\sum_{i=0}^{M} \sum_{N \in digits(i)} N(p)} \tag{3}$$

$d_k(p)$ is the effect of pixel p on model corresponds to class k, $digits(i)$ define the sample from a class of M class problem and $N(p)$ is the models output probability.

E. Region Perturbation

Region Perturbation introduced by Aopc Samek et al. [50] is an extension of Pixel-Flipping to flip an area rather than a single pixel. It has been used in several XAI experiments.

TABLE III
FAITHFULNESS CORRELATION AND FAITHFULNESS ESTIMATE METRICS IN XAI

Reference	Method	Application
[95] 2023	Guidelines for explanation evaluation	Clinical data
[42] 2023	Statistical learning for local interpretations of deep neural networks using global priors	Explanation refinement; LIME
[96] 2022	Explainability of Deep Vision-Based Autonomous Driving Systems	Autonomous driving
[97] 2022	Evaluating the Evaluation of Explainable Artificial Intelligence in Natural Language Processing	Human catered AI; Natural language understaing
[98] 2022	Explanations in Autonomous Driving	Autonomous driving
[99] 2022	Human Interpretation of Saliency-based Explanation Over Text	Human interaction; Explainability in natural language understanding
[100] 2022	Explainable predictive modelling for limited spectral data	Robustness of ML models
[101] 2022	Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning	Robustness of prediction; Faithfulness of model
[102] 2022	Information fusion as an integrative cross-cutting enabler	Legal and ethical aspect of ML; Clinical decision making
[103] 2022	Interpretability versus Explainability	Framework for interpretability and explainability
[104] 2022	Layerwise Sequential Selection (CNN) of Discernible Neurons	Understanding visual explantion
[105] 2022	On Glocal Explainability of Graph Neural Networks	Explainability; Graph neural network
[106] 2022	Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability	Attention as explanation
[107] 2022	Explainable Deep Learning: A Field Guide for the Uninitiated	Deep Learning mode Understanding
[108] 2022	Explainable deep learning in healthcare	Imterpretable deep learning in healthcare
[109] 2022	Explainable Machine Learning to Identify the Most Important Predictors of Infidelity	Personal relationship
[110] 2020	Efficient Estimation of General Additive Neural Networks	Medical decision support system
[111] 2019	Explainability in human–agent systems	General explainabiliy

TABLE IV
PIXEL FLIPPING

Reference	Method	Application
[118] 2023	Explaining the black-box smoothly	Counterfactual reasoning; Medical image understanding
[95] 2023	Post-hoc explanation from DNN	Multi-modal medical image; Post-hoc explanation
[119] 2022	Perturbation Effect	General explainability; Time series data
[120] 2022	Decoding psychophysiological EEG	Nuro-signal understanding;
[121] 2022	Spatiotemporal Prediction Model	Spatiotemporal dynamics
[122] 2022	Sensitivity of Logic Learning Machine	Autonomous driving; Feature importance
[123] 2021	Saliency by bilateral perturbations	General explainability
[124] 2021	Local Explanation Approach for Predictive Process Monitoring	Predictive process monitoring; Process mining
[125] 2020	Reliable Local Explanations	Sound analysis
[126] 2020	Interpretation by counterfactual	Medical image analysis
[127] 2019	Explanations for Attributing DNN Predictions	General XAI

Table II summarizes XAI studies, including Region Perturbation. Region perturbation metric gives Area Under Perturbation which defines by the following equation.

$$AOPC = \frac{1}{L+1}\Big\langle \sum_{k=0}^{L} f\big(x_{MoRF}^{(0)}\big) - f\big(x_{MoRF}^{(k)}\big)\Big\rangle_{p(x)} \quad (4)$$

Where f is the model, L is the number of samples, $\langle\cdot\rangle_{p(x)}$ denotes the average over all samples and $x_{MoRF}^{(k)}$ is the cumulative removal of up to k^{th} Most Relevant Feature (MoRF).

Singla et al. [118] propose a counterfactual approach to explain black-box models used for chest X-ray diagnosis. Jin et al. [14] discuss generating post-hoc explanations from deep neural networks for multi-modal medical image analysis tasks. Šimić et al. [119] introduce a perturbation effect metric to counter misleading validation of feature attribution methods in deep learning for time-series data. Huang et al. [121] focus on understanding spatiotemporal prediction models, while Narteni et al. [122] study the sensitivity of logic learning machines in safety-critical systems. Khorram et al. [123] propose an integrated gradient-optimized saliency method for explainable AI in medical imaging. In contrast, Mehdiyev and Fettke [124] provide a general overview of explainable AI for process mining with a focus on a novel local explanation approach. Mishra et al. [125] discuss reliable local explanations for machine listening, and Lenis et al. [126] introduce domain-aware medical image classifier interpretation by counterfactual impact analysis. Finally, Fong and Vedaldi [127] explain deep neural network predictions for computer vision tasks without giving detain on the evaluation of explanation. Most of the papers used pixel flipping or variants of it to evaluate the local explanations. Table IV presents the list of papers that have mentioned pixel flipping technique in their papers.

F. Selectivity

Selectivity [128] is a metrics for evaluation used in several recent XAI models, which measures how quickly a prediction function starts to drop when removing features with the highest attributed values. It can be calculated using the AOPC curve or pixel flipping curve.

G. Sensitivity-N

Sensitivity-N [47] computes the correlation between the sum of the attributions and the variation in the target output while varying the fraction of the total number of features and averages it over several test samples. This metric had been recently used by [129, 130]. For a number of features n in data, selectivity-n defines the sum of the attributions $\sum_{i=1}^{N} R_i^c(x)$ and variation in the target output correlates on a particular task for different explanation algorithms. $R_i^c(x)$ attributions of class c of input pixel i and N is the total number of pixels in the input i. gradient multiplied element-wise by the input

H. IROF

IROF introduced by Rieger and Hansen [46] computes the area over the curve per class for sorted mean importance of feature segments (superpixels) as they are iteratively removed (and prediction scores are collected), averaged over several test samples. Fel et al. [131] elaborate on the model explainability using IROF. They investigate how good the explanation is by evaluating algorithmic stability measures.

$$IROF(e_j) = \frac{1}{N}\sum_{n=1}^{N} AOC\Big(\frac{F(X_n^l)_y}{F(X_0^l)_y}\Big)_{l=0}^{L} \quad (5)$$

X_n^l denotes an augmented version of image X, with the top l of L segments replaced by their mean value due to high relevance. F represents the model, N the total test images, and AOC quantifies the area under a curve.

I. Infidelity

Infidelity is an evaluation metric introduced by [45]. It represents the expected mean square error between 1) a dot product of an attribution and input perturbation and 2) a difference in model output after significant perturbation. Lv et al. [105], Mercier et al. [132], Chatterjee et al. [133], Sahatova and Balabaeva [134], Meister et al. [135] leverage this metric in their experiments and comparisons.

$$INFD(\phi, f, x) = \mathop{\mathbb{E}}_{I\sim\mu_I}\Big[\big(I^T\phi(f,x) - (f(x) - f(x-I))\big)^2\Big] \quad (6)$$

ϕ represents the explainer, f the model, and x the input. I signifies the deviation of input from baseline x_0.

J. ROAD

ROAD (RemOve And Debias) introduced by Rong et al. [44] measures the accuracy of the model on the test set in an iterative process of removing k most important pixels, at each step k most relevant pixels (MoRF order) are replaced with noisy linear imputations. ROAD follows a similar approach to AOPC; however, the feature removal is performed using noisy approximation neighbors. To remove a pixel from an image, ROAD uses the following equation.

$$x_{i,j} = w_d(x_{i,j+1} + x_{i,j-1} + x_{i+1,j} + x_{i-1,j}) + w_i(x_{i+1,j+1} + x_{i-1,j-1} + x_{i+1,j-1} + x_{i-1,j+1}) \tag{7}$$

i, j denote pixel locations in an image. w_i and w_d are weight factors for nearest and distant neighbors respectively, with more weight given to the former in the experiment. Absent edge pixels are treated as having a value of 0.

K. Sufficiency

Sufficiency [136] measures the extent to which similar explanations have the same prediction label. For prediction explanation, if a specific property (π) justifies the prediction for an instance (x), then any other instance $(x^{'})$ with the same property (π) should also be classified similarly. In other words, consistency is required in classifying instances with the same property used for prediction justification. According to [136] to Explanations $\mathcal{E}$ are intelligible if for any instance $x \in \mathcal{X}$ and property, $\pi \in \mathcal{E}$ it is possible to assess whether π applies to x. If so, they define this as a relation $A(x^{'}, \pi)$.

$$\mathcal{C}_x = \left\{x^{'} \in \mathcal{X} : A(x^{'}, e(x))\right\} \tag{8}$$

$\mathcal{C}_x$is the set of instance that share same property as x's explanation and e is the explainer.

IV. Discussion

Applied research has seen a rise in developing and evaluating explanation evaluation metrics. While some studies use established metrics, many researchers propose their own, making it difficult to benchmark and compare methods. Furthermore, the lack of defined terminology complicates the process. However, there is potential to develop effective metrics, especially in healthcare and security domains where robust explanations are crucial [137, 138]. Establishing standard evaluation metrics is necessary to assess effectiveness and accuracy, enabling comparison and advancement in these domains.

V. conclusions

This literature review presents two taxonomies aimed at enhancing the classification of explainable AI (XAI) methods and improving the evaluation metrics used for assessing explainability in machine learning. Evaluating model explainability requires an interactive approach based on the psychological construct. Our review explores terms like interpretability and understandability in XAI evaluation. Human evaluation is prone to bias, so a formal metric that can be experimentally validated is recommended. This approach enables objective assessment and comparison of explanations across models. A formal definition of the metric will advance explainable AI and promote trustworthy machine learning systems.

References

[1] T. Miller, R. Hoffman, O. Amir, and A. Holzinger, "Special issue on explainable artificial intelligence (xai)," *Artificial Intelligence*, vol. 307, p. 103705, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370222000455

[2] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, p. 4793—4813, November 2021. [Online]. Available: https://doi.org/10.1109/TNNLS.2020.3027314

[3] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger, "Explainable ai: The new 42?" in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2018, pp. 295–303.

[4] T. Nizam and S. Zafar, *Explainable Artificial Intelligence (XAI): Conception, Visualization and Assessment Approaches Towards Amenable XAI*. Cham: Springer International Publishing, 2023, pp. 35–51. [Online]. Available: https://doi.org/10.1007/978-3-031-18292-1%5F3

[5] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research," *Artificial Intelligence*, vol. 296, p. 103473, 2021.

[6] A. Holzinger, "From machine learning to explainable ai," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018, pp. 55–66.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Inf. Fusion*, vol. 58, no. C, p. 82–115, jun 2020. [Online]. Available: https://doi.org/10.1016/j.inffus.2019.12.012

[8] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[9] J. L. Espinoza, C. L. Dupont, A. O'Rourke, S. Beyhan, P. Morales, A. Spoering, K. J. Meyer, A. P. Chan, Y. Choi, W. C. Nierman, K. Lewis, and K. E. Nelson, "Predicting antimicrobial mechanism-of-action from transcriptomes: A generalizable explainable artificial intelligence approach," *PLOS Computational Biology*, vol. 17, no. 3, pp. 1–25, 03 2021.

[10] E. Melo, I. Silva, D. G. Costa, C. M. D. Viegas, and T. M. Barros, "On the use of explainable artificial intel-

ligence to evaluate school dropout," *Education Sciences*, vol. 12, no. 12, 2022.

[11] L. Sanneman and J. A. Shah, "The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems," *International Journal of Human–Computer Interaction*, vol. 38, no. 18-20, pp. 1772–1788, 2022.

[12] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[13] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Mining and Knowledge Discovery*, Jan 2023. [Online]. Available: https://doi.org/10.1007/s10618-022-00867-8

[14] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable ai in medical image analysis," *Medical Image Analysis*, vol. 84, p. 102684, 2023.

[15] Q. V. Liao, D. Gruen, and S. Miller, "Questioning the ai: Informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1–15. [Online]. Available: https://doi.org/10.1145/3313831.3376590

[16] A. Holzinger, "From machine learning to explainable ai," in *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, 2018, pp. 55–66.

[17] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: https://doi.org/10.1145/2939672.2939778

[19] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186. [Online]. Available: https://doi.org/10.1145/3375627.3375830

[20] F. Nunnari, M. A. Kadir, and D. Sonntag, "On the overlap between grad-cam saliency maps and explainable visual features in skin cancer images," in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2021, pp. 241–253.

[21] Chinu and U. Bansal, "Explainable AI: To reveal the logic of black-box models," *New Gener. Comput.*, Feb. 2023.

[22] G. Schwalbe and B. Finzel, "A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts," *Data Min. Knowl. Discov.*, Jan. 2023.

[23] J. Xi, D. Wang, X. Yang, W. Zhang, and Q. Huang, "Cancer omic data based explainable ai drug recommendation inference: A traceability perspective for explainability," *Biomedical Signal Processing and Control*, vol. 79, p. 104144, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809422005985

[24] M. A. Kadir, A. Mohamed Selim, M. Barz, and D. Sonntag, "A user interface for explaining machine learning model explanations," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, ser. IUI '23 Companion. New York, NY, USA: Association for Computing Machinery, 2023, p. 59–63. [Online]. Available: https://doi.org/10.1145/3581754.3584131

[25] E. Melo, I. Silva, D. G. Costa, C. M. D. Viegas, and T. M. Barros, "On the use of explainable artificial intelligence to evaluate school dropout," *Educ. Sci. (Basel)*, vol. 12, no. 12, p. 845, Nov. 2022.

[26] J. Mi, L. Wang, Y. Liu, and J. Zhang, "KDE-GAN: A multimodal medical image-fusion model based on knowledge distillation and explainable AI modules," *Comput. Biol. Med.*, vol. 151, no. Pt A, p. 106273, Dec. 2022.

[27] R. Machlev, M. Perl, J. Belikov, K. Y. Levy, and Y. Levron, "Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI—explainable artificial intelligence," *IEEE Trans. Industr. Inform.*, vol. 18, no. 8, pp. 5127–5137, Aug. 2022.

[28] I. A. Khan, N. Moustafa, D. Pi, K. M. Sallam, A. Y. Zomaya, and B. Li, "A new explainable deep learning framework for cyber threat discovery in industrial IoT networks," *IEEE Internet Things J.*, vol. 9, no. 13, pp. 11 604–11 613, Jul. 2022.

[29] M. Theunissen and J. Browning, "Putting explainable AI in context: institutional explanations for medical AI," *Ethics Inf. Technol.*, vol. 24, no. 2, p. 23, May 2022.

[30] A. Ferraro, A. Galli, V. Moscato, and G. Sperlì, "Evaluating explainable artificial intelligence tools for hard disk drive predictive maintenance," *Artif. Intell. Rev.*, Dec. 2022.

[31] R. E. Sarpietro, C. Pino, S. Coffa, A. Messina, S. Palazzo, S. Battiato, C. Spampinato, and F. Rundo, "Explainable deep learning system for advanced silicon and silicon carbide electrical wafer defect map assessment," *IEEE Access*, vol. 10, pp. 99 102–99 128, 2022.

[32] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu,

C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware explainable AI," *IEEE Trans. Knowl. Data Eng.*, pp. 1–1, 2020.
[33] Y. Mualla, I. Tchappi, T. Kampik, A. Najjar, D. Calvaresi, A. Abbas-Turki, S. Galland, and C. Nicolle, "The quest of parsimonious XAI: A human-agent architecture for explanation formulation," *Artif. Intell.*, vol. 302, no. 103573, p. 103573, Jan. 2022.
[34] I. Palatnik de Sousa, M. M. B. R. Vellasco, and E. Costa da Silva, "Explainable artificial intelligence for bias detection in COVID CT-Scan classifiers," *Sensors (Basel)*, vol. 21, no. 16, p. 5657, Aug. 2021.
[35] E. Amparore, A. Perotti, and P. Bajardi, "To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods," *PeerJ Comput. Sci.*, vol. 7, no. e479, p. e479, Apr. 2021.
[36] R. R. Karn, P. Kudva, H. Huang, S. Suneja, and I. M. Elfadel, "Cryptomining detection in container clouds using system calls and explainable machine learning," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 3, pp. 674–691, Mar. 2021.
[37] S. Lobner, W. B. Tesfay, T. Nakamura, and S. Pape, "Explainable machine learning for default privacy setting prediction," *IEEE Access*, vol. 9, pp. 63 700–63 717, 2021.
[38] M. Hartmann, H. Du, N. Feldhus, I. Kruijff-Korbayová, and D. Sonntag, "XAINES: Explaining AI with narratives," *KI - Künstliche Intelligenz*, vol. 36, no. 3, pp. 287–296, Dec. 2022.
[39] M. Hartmann, I. Kruijff-Korbayová, and D. Sonntag, "Interaction with explanations in the XAINES project," *Trustworthy AI in the Wild Workshop*, vol. 9, 2021.
[40] R. Biswas, M. Barz, and D. Sonntag, "Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking," *KI - Künstl. Intell.*, vol. 34, no. 4, pp. 571–584, Dec. 2020.
[41] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Inf. Fusion*, vol. 76, pp. 89–106, Dec. 2021.
[42] X. Li, H. Xiong, X. Li, X. Zhang, J. Liu, H. Jiang, Z. Chen, and D. Dou, "G-LIME: Statistical learning for local interpretations of deep neural networks using global priors," *Artif. Intell.*, vol. 314, no. 103823, p. 103823, Jan. 2023.
[43] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, *A Benchmark for Interpretability Methods in Deep Neural Networks*. Red Hook, NY, USA: Curran Associates Inc., 2019.
[44] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "Evaluating feature attribution: An information-theoretic perspective," *CoRR*, vol. abs/2202.00449, 2022. [Online]. Available: https://arxiv.org/abs/2202.00449
[45] C.-K. Yeh, C.-Y. Hsieh, A. S. Suggala, D. I. Inouye, and P. Ravikumar, *On the (in)Fidelity and Sensitivity of Explanations*. Red Hook, NY, USA: Curran Associates Inc., 2019.
[46] L. Rieger and L. Hansen, "Irof: a low resource evaluation metric for explanation methods," in *Proceedings of the Workshop AI for Affordable Healthcare at ICLR 2020*, 2020, workshop AI for Affordable Healthcare at ICLR 2020 ; Conference date: 26-04-2020 Through 26-04-2020.
[47] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Gradient-Based attribution methods," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2019, pp. 169–191.
[48] V. Arya, R. K. E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. T. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," *CoRR*, vol. abs/1909.03012, 2019. [Online]. Available: http://arxiv.org/abs/1909.03012
[49] R. Luss, P. Chen, A. Dhurandhar, P. Sattigeri, K. Shanmugam, and C. Tu, "Generating contrastive explanations with monotonic attribute functions," *CoRR*, vol. abs/1905.12698, 2019. [Online]. Available: http://arxiv.org/abs/1905.12698
[50] W. Aopc Samek, A. Binder, G. Montavon, S. Lapuschkin, and K. R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, pp. 2660–2673, 2016.
[51] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, 07 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140
[52] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, 2021.
[53] D. Alvarez Melis and T. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf
[54] A. Kapishnikov, T. Bolukbasi, F. Viegas, and M. Terry, "Xrai: Better attributions through regions," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, nov 2019, pp. 4947–4956. [Online].

Available: https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00505

[55] A. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," *CoRR*, vol. abs/2007.07584, 2020. [Online]. Available: https://arxiv.org/abs/2007.07584

[56] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6970–6979.

[57] L. Arras, A. Osman, and W. Samek, "Clevr-xai: A benchmark dataset for the ground truth evaluation of neural network explanations," *Information Fusion*, vol. 81, pp. 14–40, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002335

[58] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 590–601.

[59] Z. Li, W. Wang, Z. Li, Y. Huang, and Y. Sato, "Towards visually explaining video understanding networks with perturbation," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1119–1128.

[60] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," *CoRR*, vol. abs/1710.11063, 2017. [Online]. Available: http://arxiv.org/abs/1710.11063

[61] S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, and J. Herlocker, "Toward harnessing user feedback for machine learning," in *Proceedings of the 12th International Conference on Intelligent User Interfaces*, ser. IUI '07. New York, NY, USA: Association for Computing Machinery, 2007, p. 82–91. [Online]. Available: https://doi.org/10.1145/1216295.1216316

[62] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local explanation methods for deep neural networks lack sensitivity to parameter values," 2018. [Online]. Available: https://openreview.net/pdf?id=SJOYTK1vM

[63] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning – a brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops*, ser. Communications in computer and information science. Cham: Springer International Publishing, 2020, pp. 417–431.

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[65] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 590–601.

[66] Z. Rguibi, A. Hajami, D. Zitouni, A. Elqaraoui, and A. Bedraoui, "Cxai: Explaining convolutional neural networks for medical imaging diagnostic," *Electronics*, vol. 11, no. 11, 2022. [Online]. Available: https://www.mdpi.com/2079-9292/11/11/1775

[67] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[68] Y. Wang, H. Su, B. Zhang, and X. Hu, "Interpret neural networks by identifying critical data routing paths," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 8906–8914. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00928

[69] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, "Understanding the decisions of cnns: An in-model approach," *Pattern Recognition Letters*, vol. 133, pp. 373–380, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865520301240

[70] M. S. Veldhuis, S. Ariëns, R. J. Ypma, T. Abeel, and C. C. Benschop, "Explainable artificial intelligence in forensics: Realistic explanations for number of contributor predictions of dna profiles," *Forensic Science International: Genetics*, vol. 56, p. 102632, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187249732100168X

[71] A. Apicella, S. Giugliano, F. Isgrò, and R. Prevete, "Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems," *Knowledge-Based Systems*, vol. 255, p. 109725, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705122008735

[72] ——, "Explanations in terms of Hierarchically organised Middle Level Features," *CEUR Workshop Proceedings*, 2021. [Online]. Available: https://ceur-ws.org/Vol-3014/paper4.pdf

[73] M. Schinle, C. Erler, M. Hess, and W. Stork, "Explainable artificial intelligence in ambulatory digital dementia screenings," *Stud. Health Technol. Inform.*, vol. 294, pp. 123–124, May 2022.

[74] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross, "A unified view of gradient-based attribution methods for deep neural networks," *CoRR*, vol. abs/1711.06104, 2017. [Online]. Available: http://arxiv.

org/abs/1711.06104
[75] P. Sturmfels, S. Lundberg, and S.-I. Lee, "Visualizing the impact of feature attribution baselines," *Distill*, 2020, https://distill.pub/2020/attribution-baselines.
[76] A. Sharma and P. K. Mishra, "Covid-MANet: Multi-task attention network for explainable diagnosis and severity assessment of COVID-19 from CXR images," *Pattern Recognit.*, vol. 131, no. 108826, p. 108826, Nov. 2022.
[77] H.-S. Kim and I. Joe, "An XAI method for convolutional neural networks in self-driving cars," *PLoS One*, vol. 17, no. 8, p. e0267282, Aug. 2022.
[78] C. Ieracitano, N. Mammone, A. Hussain, and F. C. Morabito, "A novel explainable machine learning approach for EEG-based brain-computer interface systems," *Neural Comput. Appl.*, vol. 34, no. 14, pp. 11 347–11 360, Jul. 2022.
[79] D. Lubo-Robles, D. Devegowda, V. Jayaram, H. Bedle, K. J. Marfurt, and M. J. Pranter, "Quantifying the sensitivity of seismic facies classification to seismic attribute selection: An explainable machine-learning study," *Interpretation*, vol. 10, no. 3, pp. SE41–SE69, Aug. 2022.
[80] S. Phul, G. Kuenze, C. G. Vanoye, C. R. Sanders, A. L. George, Jr, and J. Meiler, "Predicting the functional impact of KCNQ1 variants with artificial neural networks," *PLoS Comput. Biol.*, vol. 18, no. 4, p. e1010038, Apr. 2022.
[81] G. Nápoles and L. Koutsoviti Koumeri, "A fuzzy-rough uncertainty measure to discover bias encoded explicitly or implicitly in features of structured pattern classification datasets," *Pattern Recognit. Lett.*, vol. 154, pp. 29–36, Feb. 2022.
[82] A. Cartolano, A. Cuzzocrea, G. Pilato, and G. M. Grasso, "Explainable AI at work! what can it do for smart agriculture?" in *2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM)*. IEEE, Dec. 2022.
[83] M. Sabih, A. Mishra, F. Hannig, and J. Teich, "MOSP: Multi-objective sensitivity pruning of deep neural networks," in *2022 IEEE 13th International Green and Sustainable Computing Conference (IGSC)*. IEEE, Oct. 2022.
[84] G. Taskin, "A feature selection method via graph embedding and global sensitivity analysis," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[85] T. Beker, H. Ansari, S. Montazeri, Q. Song, and X. X. Zhu, "Explainability analysis of CNN in detection of volcanic deformation signal," in *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul. 2022.
[86] K. Blix, A. B. Ruescas, J. E. Johnson, and G. Camps-Valls, "Learning relevant features of optical water types," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[87] Q. Chen, G. Pan, W. Chen, and P. Wu, "A novel explainable deep belief network framework and its application for feature importance analysis," *IEEE Sens. J.*, vol. 21, no. 22, pp. 25 001–25 009, Nov. 2021.
[88] X. Wu, W. Guo, H. Wei, and X. Xing, "Adversarial policy training against deep reinforcement learning," in *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 1883–1900. [Online]. Available: https://www.usenix.org/conference/usenixsecurity21/presentation/wu-xian
[89] C. Hoyt and A. B. Owen, "Efficient estimation of the ANOVA mean dimension, with an application to neural net classification," *SIAM/ASA J. Uncertain. Quantif.*, vol. 9, no. 2, pp. 708–730, Jan. 2021.
[90] M. J. Pappaterra and F. Flammini, "Bayesian networks for online cybersecurity threat detection," in *Studies in Computational Intelligence*, ser. Studies in computational intelligence. Cham: Springer International Publishing, 2021, pp. 129–159.
[91] M. S. Kovalev and L. V. Utkin, "A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov-Smirnov bounds," *Neural Netw.*, vol. 132, pp. 1–18, Dec. 2020.
[92] K.-S. Lee, J. Y. Kim, E.-T. Jeon, W. S. Choi, N. H. Kim, and K. Y. Lee, "Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest x-ray images using explainable deep-learning algorithm," *J. Pers. Med.*, vol. 10, no. 4, p. 213, Nov. 2020.
[93] J. Kauffmann, K.-R. Müller, and G. Montavon, "Towards explaining anomalies: A deep taylor decomposition of one-class models," *Pattern Recognit.*, vol. 101, no. 107198, p. 107198, May 2020.
[94] S. M. Mathews, "Explainable artificial intelligence applications in NLP, biomedical, and malware classification: A literature review," in *Advances in Intelligent Systems and Computing*, ser. Advances in intelligent systems and computing. Cham: Springer International Publishing, 2019, pp. 1269–1292.
[95] W. Jin, X. Li, M. Fatehi, and G. Hamarneh, "Guidelines and evaluation of clinical explainable AI in medical image analysis," *Med. Image Anal.*, vol. 84, no. 102684, p. 102684, Feb. 2023.
[96] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems: Review and challenges," *Int. J. Comput. Vis.*, vol. 130, no. 10, pp. 2425–2452, Oct. 2022.
[97] M. Neely, S. F. Schouten, M. Bleeker, and A. Lucic, "A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing," in *HHAI2022: Augmenting Human Intellect*, ser. Frontiers in artificial intelligence and applications. IOS Press, Sep. 2022.
[98] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 10 142–10 162, Aug. 2022.

[99] H. Schuff, A. Jacovi, H. Adel, Y. Goldberg, and N. T. Vu, “Human interpretation of saliency-based explanation over text,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jun. 2022.

[100] F. Akulich, H. Anahideh, M. Sheyyab, and D. Ambre, “Explainable predictive modeling for limited spectral data,” *Chemometr. Intell. Lab. Syst.*, vol. 225, no. 104572, p. 104572, Jun. 2022.

[101] W. Zhang, M. Dimiccoli, and B. Y. Lim, “Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning,” in *CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, Apr. 2022.

[102] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. D. Ser, W. Samek, I. Jurisica, and N. Díaz-Rodríguez, “Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence,” *Inf. Fusion*, vol. 79, pp. 263–278, Mar. 2022.

[103] I. Namatēvs, K. Sudars, and A. Dobrājs, “Interpretability versus explainability: Classification for understanding deep learning systems and models,” *Computer Assisted Methods in Engineering and Science*, vol. 29, no. 4, pp. 297–356, 2022. [Online]. Available: https://cames.ippt.pan.pl/index.php/cames/article/view/518

[104] M. T. B. Iqbal, A. Muqeet, and S.-H. Bae, “Visual interpretation of CNN prediction through layerwise sequential selection of discernible neurons,” *IEEE Access*, vol. 10, pp. 81 988–82 002, 2022.

[105] G. Lv, L. Chen, and C. C. Cao, “On glocal explainability of graph neural networks,” in *Database Systems for Advanced Applications*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2022, pp. 648–664.

[106] M. Tutek and J. Snajder, “Toward practical usage of the attention mechanism as a tool for interpretability,” *IEEE Access*, vol. 10, pp. 47 011–47 030, 2022.

[107] G. Ras, N. Xie, M. Van Gerven, and D. Doran, “Explainable deep learning: A field guide for the uninitiated,” *J. Artif. Intell. Res.*, vol. 73, pp. 329–397, Jan. 2022.

[108] D. Jin, E. Sergeeva, W.-H. Weng, G. Chauhan, and P. Szolovits, “Explainable deep learning in healthcare: A methodological survey from an attribution view,” *WIREs Mech. Dis.*, vol. 14, no. 3, p. e1548, May 2022.

[109] L. M. Vowels, M. J. Vowels, and K. P. Mark, “Is infidelity predictable? using explainable machine learning to identify the most important predictors of infidelity,” *J. Sex Res.*, vol. 59, no. 2, pp. 224–237, Feb. 2022.

[110] P. J. G. Lisboa, S. Ortega-Martorell, M. Jayabalan, and I. Olier, “Efficient estimation of general additive neural networks: A case study for CTG data,” in *ECML PKDD 2020 Workshops*, ser. Communications in computer and information science. Cham: Springer International Publishing, 2020, pp. 432–446.

[111] A. Rosenfeld and A. Richardson, “Explainability in human–agent systems,” *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, p. 673–705, nov 2019. [Online]. Available: https://doi.org/10.1007/s10458-019-09408-y

[112] B. X. Yong and A. Brintrup, “Coalitional bayesian autoencoders: Towards explainable unsupervised deep learning with applications to condition monitoring under covariate shift,” *Appl. Soft Comput.*, vol. 123, no. 108912, p. 108912, Jul. 2022.

[113] E. Albini, J. Long, D. Dervovic, and D. Magazzeni, “Counterfactual shapley additive explanations,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: ACM, Jun. 2022.

[114] M. L. Baptista, K. Goebel, and E. M. P. Henriques, “Relation between prognostics predictor evaluation metrics and local interpretability SHAP values,” *Artif. Intell.*, vol. 306, no. 103667, p. 103667, May 2022.

[115] A. Fernandez, F. Herrera, O. Cordon, M. Jose del Jesus, and F. Marcelloni, “Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?” *IEEE Comput. Intell. Mag.*, vol. 14, no. 1, pp. 69–81, Feb. 2019.

[116] A. Wullenweber, A. Akman, and B. W. Schuller, “CoughLIME: Sonified explanations for the predictions of COVID-19 cough classifiers,” in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, Jul. 2022.

[117] V. Pitroda, M. M. Fouda, and Z. M. Fadlullah, “An explainable AI model for interpretable lung disease classification (2021),” in *Proceedings of the 2021 IEEE International Conference on Internet of Things and Intelligence Systems*, 2021, pp. 98–103.

[118] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, “Explaining the black-box smoothly-a counterfactual approach,” *Med. Image Anal.*, vol. 84, no. 102721, p. 102721, Feb. 2023.

[119] I. Šimić, V. Sabol, and E. Veas, “Perturbation effect,” in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. New York, NY, USA: ACM, Oct. 2022.

[120] J. Żygierewicz, R. A. Janik, I. T. Podolak, A. Drozd, U. Malinowska, M. Poziomska, J. Wojciechowski, P. Ogniewski, P. Niedbalski, I. Terczynska, and J. Rogala, “Decoding working memory-related information from repeated psychophysiological EEG experiments using convolutional and contrastive neural networks,” *J. Neural Eng.*, vol. 19, no. 4, p. 046053, Sep. 2022.

[121] X. Huang, X. Li, Y. Ye, S. Feng, C. Luo, and B. Zhang, “On understanding of spatiotemporal prediction model,” *IEEE Trans. Circuits Syst. Video Technol.*, pp. 1–1, 2022.

[122] S. Narteni, V. Orani, I. Vaccari, E. Cambiaso, and M. Mongelli, “Sensitivity of logic learning machine for reliability in safety-critical systems,” *IEEE Intell. Syst.*,

vol. 37, no. 5, pp. 66–74, Sep. 2022.

[123] S. Khorram, T. Lawson, and L. Fuxin, "iGOS++," in *Proceedings of the Conference on Health, Inference, and Learning*. New York, NY, USA: ACM, Apr. 2021.

[124] N. Mehdiyev and P. Fettke, "Explainable artificial intelligence for process mining: A general overview and application of a novel local explanation approach for predictive process monitoring," in *Studies in Computational Intelligence*, ser. Studies in computational intelligence. Cham: Springer International Publishing, 2021, pp. 1–28.

[125] S. Mishra, E. Benetos, B. L. T. Sturm, and S. Dixon, "Reliable local explanations for machine listening," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Jul. 2020.

[126] D. Lenis, D. Major, M. Wimmer, A. Berg, G. Sluiter, and K. Bühler, "Domain aware medical image classifier interpretation by counterfactual impact analysis," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2020, pp. 315–325.

[127] R. Fong and A. Vedaldi, "Explanations for attributing deep neural network predictions," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2019, pp. 149–167.

[128] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200417302385

[129] G. Jeon, H. Jeong, and J. Choi, "Distilled gradient aggregation: Purify features for input attribution in the deep neural network," in *Advances in Neural Information Processing Systems*, 2022.

[130] M. Ancona, *Attribution Methods for Interpreting and Optimizing Deep Neural Networks (Doctoral dissertation)*. Zurich, Switzerland: ETH Zurich, 2020.

[131] T. Fel, D. Vigouroux, R. Cadene, and T. Serre, "How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2022.

[132] D. Mercier, A. Dengel, and S. Ahmed, "TimeREISE: Time series randomized evolving input sample explanation," *Sensors (Basel)*, vol. 22, no. 11, p. 4084, May 2022.

[133] S. Chatterjee, A. Das, C. Mandal, B. Mukhopadhyay, M. Vipinraj, A. Shukla, R. Nagaraja Rao, C. Sarasaen, O. Speck, and A. Nürnberger, "TorchEsegeta: Framework for interpretability and explainability of image-based deep learning models," *Appl. Sci. (Basel)*, vol. 12, no. 4, p. 1834, Feb. 2022.

[134] K. Sahatova and K. Balabaeva, "An overview and comparison of XAI methods for object detection in computer tomography," *Procedia Comput. Sci.*, vol. 212, pp. 209–219, 2022.

[135] S. Meister, M. Wermes, J. Stüve, and R. M. Groves, "Investigations on explainable artificial intelligence methods for the deep learning classification of fibre layup defect in the automated composite manufacturing," *Compos. B Eng.*, vol. 224, no. 109160, p. 109160, Nov. 2021.

[136] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for evaluating faithfulness of local explanations," *CoRR*, Feb. 2022.

[137] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2018, pp. 0210–0215.

[138] T. Ploug and S. Holm, "The four dimensions of contestable ai diagnostics - a patient-centric approach to explainable ai," *Artificial Intelligence in Medicine*, vol. 107, p. 101901, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365720301330