Assessing XAI: Unveiling Evaluation Metrics for Local Explanation, Taxonomies, Key Concepts, 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]) is concerned with the development of methodologies for understanding the underlying logic behind machine learning and deep learning models' decision and generally, widely AI applications [2, 3]. The ultimate aspiration of XAI is to facilitate users in constructing a comprehensive and accurate cognitive representation of machine learning algorithm to encourage confidence in its outputs [4, 5, 6]. Despite the numerous methods and approaches proposed for explainability, scholars still have no agreement on what an Explanation truly is and what properties should be considered for it to be practical and understandable for the end user [7]. Future research should focus on the definition of explainability and structured formats of Explanations that can encompass various attributes and dimensions, incorporating as many aspects as possible.

Additionally, explainability should be considered a concept from psychology. It is linked to constructs such as trust, transparency, and privacy, and humans are the final consumers of explanations[8]. It reminds us that interactive visual explanations are not an optional aspect of XAI. Thus, in order to develop successful XAI models, research on psychometrics should be performed. Many authors studied the explainability in AI sub-domains through case studies or surveys, motivating the need for fundamental research on measuring explainability[9, 10, 11].

Researchers reviewed methods for explanations with neural and Bayesian networks and clustered scientific contributions devoted to extracting rules. The goals are to create rules interpretable by humans while maintaining accuracy [12]. Few scholars attempted a comprehensive survey and organization of explainability methods as a whole [13]. Others identified a broad set of requirements an explanation should meet to be understandable by laypeople and provide actionable information to support decision-making [14, 15, 16]. In their works, a thorough evaluation analysis highlighted the need for a systematic analysis of the metrics related to explainability.

Consequently, this paper aims to fill this gap by systematically reviewing research studies in XAI, focusing on a subset of peer-reviewed articles that tackled explainability from a conceptual and theoretical point of view and proposed approaches to evaluate XAI methods. The conceptual framework is that explainers build explainability methods that can be evaluated using evaluation metrics. When creating interactive visual explanations, explainers often rely on complex models or numerical approximations [17, 18, 19, 20].

However, without a robust evaluation metric for explanations, it is not easy to generalize the causality between a trained model and its visual explainer. The absence of robust evaluation can lead to inconsistencies and inaccuracies in the

Identify applicable funding agency here. If none, delete this.

explanations provided, making it challenging for users to understand the underlying concepts fully. Therefore, developing a reliable evaluation metric is essential to ensure interactive visual explanations are compelling and informative. With such a metric, explanation quality and reliability can be better assessed, allowing them to create more accurate and valuable visualizations for AI practitioners.

II. Methodology

The methodology is based on the comprehensive literature search integrated with systematic screening and literature review following the PRISMA guideline. The fundamental search database is Scopus[1] complemented with Google Scholar [2]. Scopus is a reliable search engine for peer-reviewed scientific literature. Scopus indexes journals and conference proceedings articles that meet adequate scientific standards and measures. For this research, Scopus has been used to find relevant articles. However, pioneers of AI may also use Arxiv[3] and other preprints to share an early version of their articles to make the results available faster. We also explore Google Scholar when Scopus fails to find adequate literature or when complementary information is needed. The initial queries are 'explainable artificial intelligence', 'XAI', 'explainable AI', 'explainable machine learning', and 'explainable deep learning'. The queries result in 6122 articles, including various methods and applications of explainable machine learning and deep learning.

The research methodology flow diagram, which follows the PRISMA[4] guideline, is visualized in Fig. 1. Through the methodology ' fundamental three phases, the initial 6122 documents are refined to the desired and relevant articles, including the XAI evaluation metrics. The initial documents include a vast number of documents written about XAI. Further screening is essential to refine what articles include explainability evaluation metrics. The challenge is that there is no common keyword to look for as explainability in various applications and scientific domains might be defined differently. We included a comprehensive list of keywords and relevant phrases generally used to define explainability. Here, it is worth mentioning that relying on the keywords of 'evaluation metrics' or 'metrics' cannot be efficient in finding the relevant articles for explainability evaluation metrics because explainability can be defined and communicated through different phrases, and also, the term 'evaluation metrics' might not be directly communicated throughout the literature. Therefore, exploring the literature using the relevant keywords to explainability complemented with the keywords of 'evaluation metrics' or 'metrics' is considered an efficient method to remove the 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.

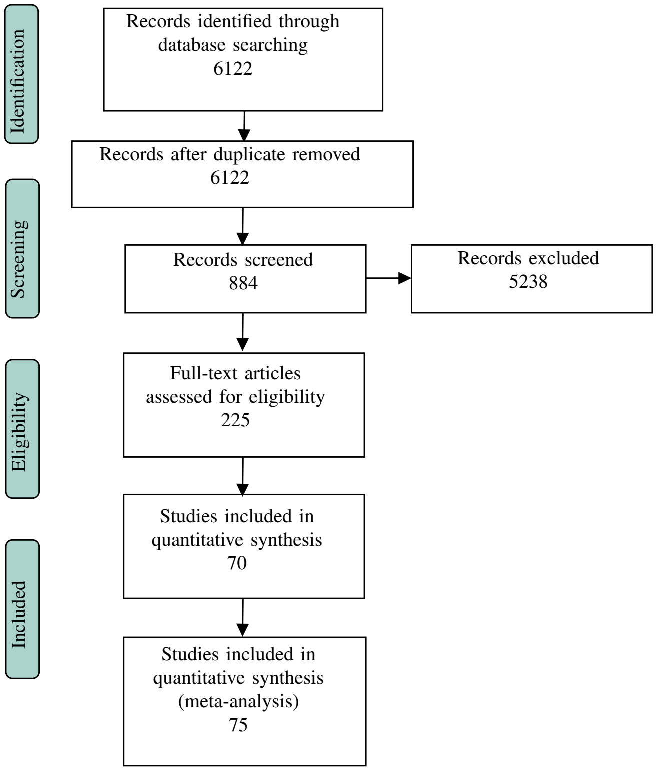
Identification
Screening
Eligibility
Included
Records identified through database searching 6122
Records after duplicate removed 6122
Records screened 884
Records excluded 5238
Full-text articles assessed for eligibility 225
Studies included in quantitative synthesis 70
Studies included in quantitative synthesis (meta-analysis) 75

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.

After a second screening by reading and going through articles the 155 articles had been removed. The excluded articles were concerned with the evaluation metrics irrelevant to explainability. Often the performance and accuracy had been measured and the explainability had been discussed shortly. In the final stage the articles were reduced to 75. At the next stage the original studies and most relevant articles are selected for review resulting in 70 articles classified according to the metrics presented in the tables.

III. Result

XAI research has exponentially expanded during the past five years. Fig. 2 on the next page illustrates the research progress based on the number of published articles. The initial inquiry is to find articles devoted to XAI results in 6122 articles. Explainability metrics are used to evaluate the quality of the explainer or model for investigating how good the explanation is. A limited number of articles in the literature include evaluation metrics for measuring explainability. Around 884

[1] https://www.scopus.com/home.uri

[2] https://scholar.google.com

[3] https://arxiv.org

[4] http://www.prisma-statement.org/

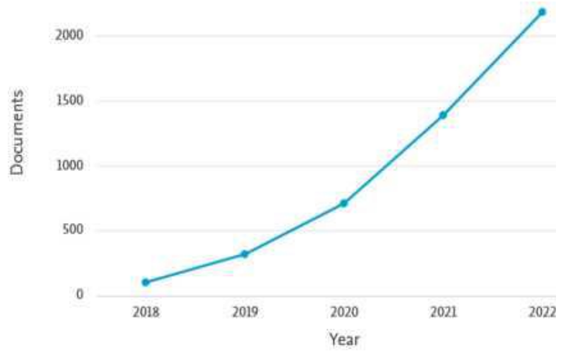

Fig. 2. The initial queries for XAI literature resulted in 6122 articles. The graph also indicates an upward trend in XAI research.

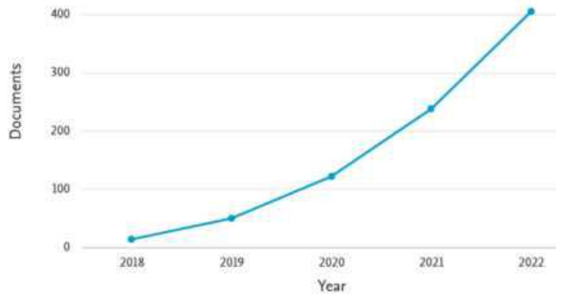

Fig. 3. After the first screening for explainability evaluation metrics, the literature review revealed that there were less than half a thousand research articles containing explanation metrics in 2022. However, there is an upward trend in the use of explanation metrics in research.

articles include the evaluation inquiry in the title, abstract, or keywords. Fig. 3 illustrates the distribution of other results during the past five years.

Table III on the next page includes many studies on Explainable Artificial Intelligence (XAI) written between 2021 and 2023. The authors have looked into various XAI-related ideas, methods, and applications. The papers span multiple topics, including sentiment analysis, cybersecurity, healthcare, education, and industry. The authors used deep learning, knowledge distillation, statistical testing, and other techniques to create more accessible models for people to understand and use. They have suggested assessment techniques, metrics, and algorithms to measure how trustworthy and comprehensible AI models are. To increase the interpretability of black-box models, the authors have also leveraged various XAI tools, including SHAP, LIME, and LEAF. In general, the contributions of the authors seek to improve human-AI collaboration and address the difficulties of implementing AI systems. As an AI application becomes ubiquitous, the demand for explainability also grows. To address this, researchers have proposed different evaluation metrics to measure the quality of explanations produced by AI models.

According to Chinu and Bansal [21], the explanation metric is crucial for assessing the effectiveness of answers in relief application submissions. According to a study of Schwalbe and Finzel [13], applications based on explainable techniques and their matrices are gaining popularity. For example, Machlev et al. [27] emphasizes the significance of explainability metrics for power experts in understanding power quality distribution classification to guarantee reliable decision-making. The importance of explanation and its evaluation metric in locating anomalies in IoT data for enhancing security was stated by [28]. Vilone and Longo [41] discovered that despite the substantial body of knowledge created regarding explainability, academics still need to have a general agreement regarding how to define an explanation, and its validity and reliability can be evaluated. In addition, Li et al. [32] note that although many works have contributed to this line, the present endeavor lacks a clear taxonomy and systematic review. To address this, Mualla et al. [33], Li et al. [42] propose new explanation techniques and use the metric used in LIME to evaluate the quality of the explanation.

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 Fig. III on page 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

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.

Authors	Year	Source title	Concept	Application
Chinu and Bansal [21]	2023	New Generation Computing	Explainable AI: To Reveal the Logic of Black-Box Models	Interpretable; Transparency; Quality metrics;
Schwalbe and Finzel [22]	2023	Data Mining and Knowledge Discovery	A taxonomy for XAI methods	XAI; Interpretability; Meta-analysis
Xi et al. [23]	2023	Biomedical Signal Processing and Control	XAI Evaluation metric: Traceability rate	Drug recommendation; Explainability; Traceability
Kadir et al. [24]	2023	IUI'23 Companion	Explaining Machine Learning Model Explanations	Interpretability; GUI for Explanation
Melo et al. [25]	2022	Education Sciences	XAI methods evaluation metric	Educational data science; Learning analytics
Mi et al. [26]	2022	Computers in Biology and Medicine	KDE-GAN: A multimodal medical image-fusion model knowledge distillation and explainable AI	XAI in Medicine; Image generation
Machlev et al. [27]	2022	IEEE Transactions on Industrial Informatics	Measuring Explainability and Trustworthiness of Power Quality Disturbances Classifiers	XAI in Power; Power quality disturbances (PQDs)
Khan et al. [28]	2022	IEEE Internet of Things Journal	A New Explainable Deep Learning Framework for Cyber Threat Discovery	Anomaly detection; IIoT; industrial networks
Theunissen and Browning [29]	2022	Ethics and Information Technology	Putting explainable AI in context: institutional explanations for medical AI	AI and health; Epistemic risk; Ethical design
Ferraro et al. [30]	2022	Artificial Intelligence Review	Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance	Predictive maintenance
Sarpietro et al. [31]	2022	IEEE Access	Explainable Deep Learning System for Advanced Silicon and Silicon Carbide Electrical Wafer Defect Map Assessment	Explainable architectures; Hierarchical clustering
Li et al. [32]	2020	IEEE Transactions on Knowledge and Data Engineering	A Survey of Data-Driven and Knowledge-Aware eXplainable AI	Knowledge-base;
Mualla et al. [33]	2022	Artificial Intelligence	A human-agent architecture for explanation formulation	HCI; Multi-agent systems
Vilone and Longo [12]	2021	Information Fusion	Notions of explainability and evaluation approaches for explainable artificial intelligence	Evaluation methods; Notions of explainability
Palatnik de Sousa et al. [34]	2021	Sensors	Explainable artificial intelligence for bias detection	Computerised Tomography
Amparore et al. [35]	2021	PeerJ Computer Science	To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods	Local linear explanation; Machine Learning Auditing
Karn et al. [36]	2021	IEEE Transactions on Parallel and Distributed Systems	Cryptomining Detection in Container Clouds Using System Calls and Explainable Machine Learning	Anomaly detection; Cryptomining; explainability
Lobner et al. [37]	2021	IEEE Access	Explainable Machine Learning for Default Privacy Setting Prediction	Privacy preference
Hartmann et al. [38]	2022	Springer link	Explaining AI with Narratives	Explainability, NL
Hartmann et al. [39]	2021	Trustworthy AI in the wild	Interaction with Explanations	User interaction
Hartmann et al. [38]	2022	ACL	A survey on improving NLP models with human explanations	User interaction
Biswas et al. [40]	2020	Springer Link	Explanatory Interactive Image Captioning	Image captioning

the explanation's reliability and usability.

An alternative taxonomy is proposed in 5. In our recent review of application-related research, we have identified that the evaluation technique is not the sole focus of interest but

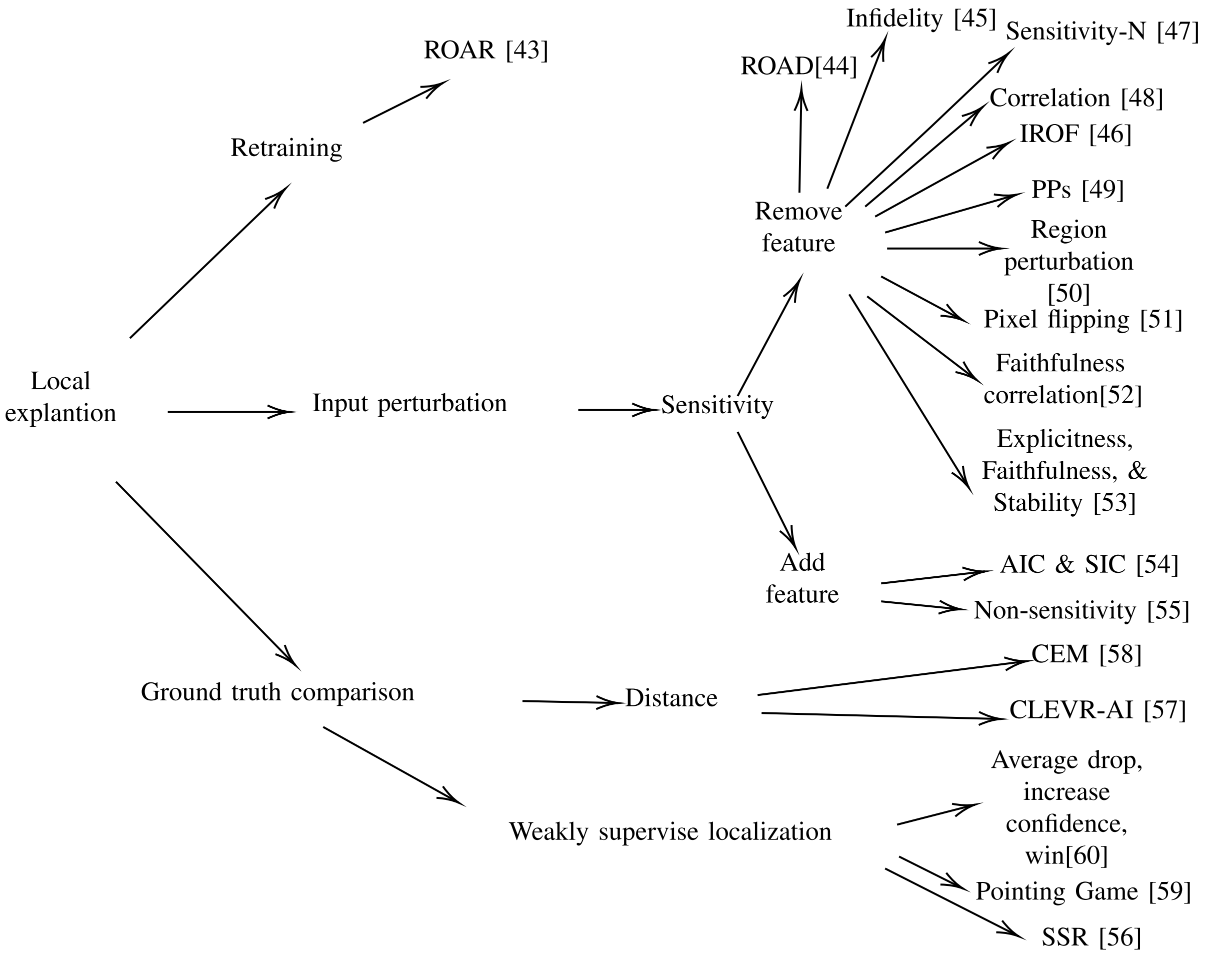

Fig. 4. Proposed taxonomy based the methodologies of the explainability evaluation

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 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 III on the previous page, 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-

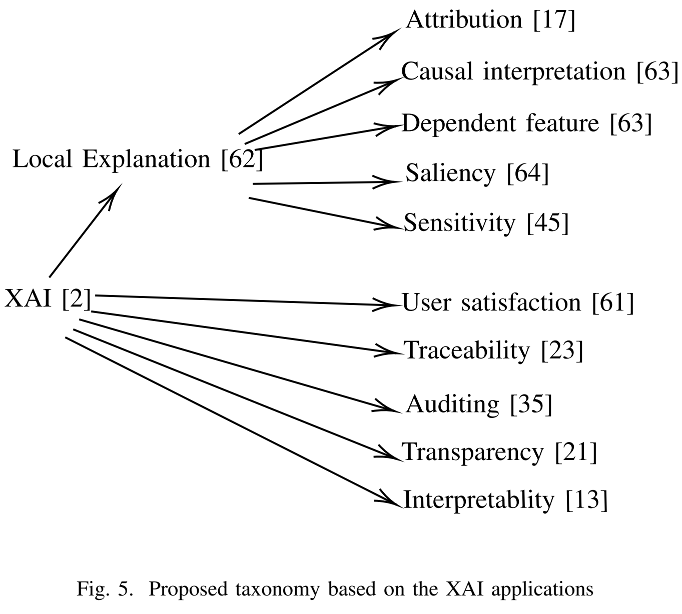

Fig. 5. Proposed taxonomy based on the XAI applications

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 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 futures 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 on the following page 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 $j \in \{1, ..., d\}$,

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

where $e_j \in R_d$ is the j^{th} coordinate basis vector, with j^{th} entry one and all others zero. 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 and e_j is the changes in the input features.

Table II on the next page 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 probability drops and attribution scores on various points. Table III on page 9 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:

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

TABLE II
SENSITIVITY ANALYSIS FOR XAI METHODS

Reference	Year	Source title	XAI method	Application
Sharma and Mishra [76]	2022	Pattern Recognition	Covid-MANet	Sensitivity analysis; Lesion localisation

Kim and Joe [77]	2022	PLoS ONE	An XAI method for convolutional neural networks	Self driving car;
				CNN;
				Sensitivity of features

Ieracitano et al. [78]	2022	Neural Computing and Applications	A novel explainable machine learning approach for EEG-based brain-computer interface systems	Brain–computer interface; Explainable machine learning
Lubo-Robles et al. [79]	2022	Interpretation	Quantifying the sensitivity of seismic facies classification to seismic attribute selection	Sensitivity of seismic attributes; Seismic geomorphology
Phul et al. [80]	2022	PLoS Computational Biology	Predicting the functional impact of KCNQ1 variants with artificial neural networks	Protein structure
Nápoles and Koutsoviti Koumeri [81]	2022	Pattern Recognition Letters	Discover bias in features of structured pattern classification datasets	Understanding bias; Fairness
Cartolano et al. [82]	2022	IEEE 8th International Conference on Multimedia Big Data	Explainable AI at Work! What Can It Do for Smart Agriculture?	Explainability in Agriculture data
Sabih et al. [83]	2022	IEEE International Green and Sustainable Computing	MOSP: Pruning of Deep Neural Networks	Neural network compression
Taskin [84]	2022	IEEE Geoscience and Remote Sensing Letters	A Feature Selection Method via Graph Embedding and Global Sensitivity Analysis	Feature engineering
Beker et al. [85]	2022	International Geoscience and Remote Sensing Symposium	Explainability Analysis of CNN in Detection of Volcanic Deformation Signal	Volcanic Deformation analysis
Blix et al. [86]	2022	IEEE Geoscience and Remote Sensing Letters	Learning Relevant Features of Optical Water Types	Understanding See water
Chen et al. [87]	2021	IEEE Sensors Journal	Deep belief network framework and its application for feature importance analysis	Feature engineering
Apicella et al. [72]	2021	CEUR Workshop Proceedings	Explanations in terms of Hierarchically organised Middle Level Features	Feature understanding
Wu et al. [88]	2021	USENIX Security Symposium	Adversarial policy training against deep reinforcement learning	Preventing adversarial attacks
Hoyt and Owen [89]	2021	SIAM-ASA Journal on Uncertainty Quantification	Efficient Estimation of the ANOVA Mean Dimension, with an Application to Neural Net Classification	Dimensionality reduction
Pappaterra and Flammini [90]	2021	Studies in Computational Intelligence	Bayesian Networks for Online Cybersecurity Threat Detection	Threat detection and analysis
Kovalev and Utkin [91]	2020	Neural Networks	A robust algorithm for explaining unreliable machine learning survival models	Reducing data demand
Lee et al. [92]	2020	Journal of Personalized Medicine	Evaluation of scalability and degree of fine-tuning	Medical imaging; Low training data
Kauffmann et al. [93]	2020	Pattern Recognition	A deep Taylor decomposition of one-class models	Outlier detection; Unsupervised learning
Molnar et al. [63]	2020	Communications in Computer and Information Science	Interpretable Machine Learning A Brief History	Dependent features; Causal interpretation
Mathews [94]	2019	Advances in Intelligent Systems and Computing	XAI Applications in NLP; Biomedical Classification	Drawback of blackbox model

d is the dimension of x.

Table III on the next page 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. $N(p)$ 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 3 summarises XAI studies, including Region Perturbation. Region perturbation metric gives Area Under Perturbation which defines by the following equation.

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

Where f is tthe 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. [123] propose an integrated gradient-optimized saliency method for explainable AI in medical imaging. In contrast, [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 on page 10 presents the list of papers that have mentioned pixel flipping technique in their papers.

F. Selectivity

Selectivity [128] in 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. Goswami et al. [129], Vangala et al. [130], Kim and Park [131], Feldmann and Bajorath [132], Wang et al. [133]. 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 [134, 135]. 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.

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. [136] elaborate on the model explainability using IROF. They investigate how good the explanation is by evaluating algorithmic stability measures.

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

I. Infidelity

Infidelity is an evaluation metric introduced by [45]. It represents the expected mean square error between 1) a dot

TABLE III
FAITHFULNESS CORRELATION AND FAITHFULNESS ESTIMATE METRICS IN XAI

References	Year	Source title	XAI method	Application
Jin et al. [95]	2023	Medical Image Analysis	Guidelines and evaluation of clinical explainable AI in medical image analysis	Guidelines for explanation evaluation; Clinical data
Li et al. [42]	2023	Artificial Intelligence	Statistical learning for local interpretations of deep neural networks using global priors	Explanation refinement; LIME
Zablocki et al. [96]	2022	International Journal of Computer Vision	Explainability of Deep Vision-Based Autonomous Driving Systems	Autonomous driving
Neely et al. [97]	2022	Frontiers in Artificial Intelligence and Applications	Evaluating the Evaluation of Explainable Artificial Intelligence in Natural Language Processing	Human catered AI; Natural language understaing
Omeiza et al. [98]	2022	IEEE Transactions on Intelligent Transportation Systems	Explanations in Autonomous Driving	Autonomous driving
Schuff et al. [99]	2022	ACM International Conference Proceeding Series	Human Interpretation of Saliency-based Explanation Over Text	Human interaction; Explainability in natural language understanding
Akulich et al. [100]	2022	Chemometrics and Intelligent Laboratory Systems	Explainable predictive modelling for limited spectral data	Robustness of ML models
Zhang et al. [101]	2022	Conference on Human Factors in Computing Systems - Proceedings	Debiased-CAM to mitigate image perturbations with faithful visual explanations of machine learning	Robustness of prediction; Faithfulness of model
Holzinger et al. [102]	2022	Information Fusion	Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and rustworthy medical artificial intelligence	Legal and ethical aspect of ML; Clinical decision making
Namatēvs et al. [103]	2022	Computer Assisted Methods in Engineering and Science	Interpretability versus Explainability	Framework for interpretability and explainability
Iqbal et al. [104]	2022	IEEE Access	Visual Interpretation of CNN Prediction Through Layerwise Sequential Selection of Discernible Neurons	Understanding visual explantion
Lv et al. [105]	2022	Lecture Notes in Computer Science	On Glocal Explainability of Graph Neural Networks	Explainability; Graph neural network
Tutek and Snajder [106]	2022	IEEE Access	Toward Practical Usage of the Attention Mechanism as a Tool for Interpretability	Attention as explanation
Ras et al. [107]	2022	Journal of Artificial Intelligence Research	Explainable Deep Learning: A Field Guide for the Uninitiated	Deep Learning mode Understanding
Jin et al. [108]	2022	WIREs Mechanisms of Disease	Explainable deep learning in healthcare	Imterpretable deep learning in healthcare
Vowels et al. [109]	2022	Journal of Sex Research	Explainable Machine Learning to Identify the Most Important Predictors of Infidelity	Personal relationship
Lisboa et al. [110]	2020	Communications in Computer and Information Science	Efficient Estimation of General Additive Neural Networks	Medical decision support system
Rosenfeld and Richardson [111]	2019	Autonomous Agents and Multi-Agent Systems	Explainability in human–agent systems	General explainabiliy

TABLE IV
PIXEL FLIPPING

Authors	Year	Source title	Title	Application
Singla et al. [118]	2023	Medical Image Analysis	Explaining the black-box smoothly	Counterfactual reasoning; Medical image understanding
Jin et al. [95]	2023	MethodsX	Generating post-hoc explanation from deep neural networks for multi-modal tasks	Multi-modal medical image; Post-hoc explanation
Šimić et al. [119]	2022	International Conference on Information and Knowledge Management, Proceedings	Perturbation Effect	General explainability; Time series data
Żygierewicz et al. [120]	2022	Journal of Neural Engineering	Decoding working memory-related information from repeated psychophysiological EEG	Nuro-signal understanding;
Huang et al. [121]	2022	IEEE Transactions on Circuits and Systems for Video Technology	On Understanding of Spatiotemporal Prediction Model	Spatiotemporal dynamics
Narteni et al. [122]	2022	IEEE Intelligent Systems	Sensitivity of Logic Learning Machine for Reliability in Safety-Critical Systems	Autonomous driving; Feature importance
Khorram et al. [123]	2021	ACM Conference on Health, Inference, and Learning	IGOS++: Integrated gradient optimized saliency by bilateral perturbations	General explainability
Mehdiyev and Fettke [124]	2021	Studies in Computational Intelligence	Application of a Novel Local Explanation Approach for Predictive Process Monitoring	Predictive process monitoring; Process mining
Mishra et al. [125]	2020	Proceedings of the International Joint Conference on Neural Networks	Reliable Local Explanations for Machine Listening	Sound analysis
Lenis et al. [126]	2020	Lecture Notes in Computer Science	Domain aware medical image classifier interpretation by counterfactual impact analysis	Medical image analysis
Fong and Vedaldi [127]	2019	Lecture Notes in Computer Science	Explanations for Attributing Deep Neural Network Predictions	General XAI

product of an attribution and input perturbation and 2) a difference in model output after significant perturbation. Lv et al. [105], Mercier et al. [137], Chatterjee et al. [138], Sahatova and Balabaeva [139], Meister et al. [140] leverage this metric in their experiments and comparisons.

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

ϕ is the explainer, f is the model x is the input $I = x - x_0$is the difference between input and baseline.

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.

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

w_i and w_d are two different weight factor for nearest and distant neighbour pixels. In the experiment they use larger weight for nearest neighbour than the distant neighbour.

K. Sufficiency

Sufficiency [141] measures the extent to which similar explanations have the same prediction label. To explain a prediction, it is necessary that if a certain property (π) is used to justify the prediction of an instance (x). Any other instance $(x^{'})$ with the same property (π) should also be classified in the same way. In other words, consistency is required in classifying instances with the same property used for prediction justification. According to [141] 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 significantly increased, focusing on developing and evaluating explanation evaluation metrics. While some studies use established metrics to measure the performance of their explanation methods, many researchers have proposed their metrics. This trend has made benchmarking and comparing different explanation methods challenging because they are evaluated using different metrics. Additionally, some terminology related to explanation evaluation still needs to be defined, further complicating the process.

However, despite these challenges, there is a significant potential for developing effective explanation evaluation metrics, particularly in domains such as healthcare and security. As these domains involve sensitive and critical decision-making processes, having robust and reliable explanation methods is crucial [142, 143]. Therefore, there is a need to establish standard evaluation metrics that can be used to assess the effectiveness and accuracy of explanation methods. Standard evaluation metrics will help compare different methods and help develop more effective and trustworthy explanation methods for these domains.

V. CONCLUSIONS

This study presented a comprehensive literature review on evaluation metrics for explainability. Two taxonomy has been proposed to bring insight into applications and evaluation of XAI. Evaluating the explainability of the models consists of an interactive approach that relays on an understanding of the psychological construct of explainability. Our review explored explainability, and terms like interpretability and understandability were explored to bring insight into XAI evaluation and metrics. Explainability has been discussed in many different areas and should be further explored to create a definition that can be used in various contexts.

To address the challenges in evaluating the quality of explanations generated by XAI methods, we propose a theoretically robust metric that can be generalized to any explanation algorithm. Our review of existing methods reveals that relying on human users to evaluate the quality of explanations can be error-prone due to the risk of confirmation bias. Therefore, we suggest defining a formal explanation evaluation metric that can be experimentally validated based on established methods. By doing so, we can ensure that the quality of explanations is objectively assessed and compared across different models and algorithms. A formal definition explanation evaluation metric will help advance the field of explainable AI and promote the development of trustworthy and transparent machine learning systems.

Acknowledgments 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.

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 intelligence 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] P. P. Goswami, T. Deshpande, D. R. Rotake, and S. G. Singh, "Near perfect classification of cardiac biomarker Troponin-I in human serum assisted by SnS2-CNT composite, explainable ML, and operating-voltage-selection-algorithm," *Biosens. Bioelectron.*, vol. 220, no. 114915, p. 114915, Jan. 2023.

[130] S. R. Vangala, N. Bung, S. R. Krishnan, and A. Roy, "An interpretable machine learning model for selectivity of small-molecules against homologous protein family," *Future Med. Chem.*, vol. 14, no. 20, pp. 1441–1453, Oct. 2022.

[131] I. B. Kim and S.-C. Park, "Machine learning-based definition of symptom clusters and selection of antidepressants for depressive syndrome," *Diagnostics (Basel)*, vol. 11, no. 9, p. 1631, Sep. 2021.

[132] C. Feldmann and J. Bajorath, "Differentiating inhibitors of closely related protein kinases with single- or multi-target activity via explainable machine learning and feature analysis," *Biomolecules*, vol. 12, no. 4, p. 557, Apr. 2022.

[133] Y. Wang, M. Huang, H. Deng, W. Li, Z. Wu, Y. Tang, and G. Liu, "Identification of vital chemical information via visualization of graph neural networks," *Brief. Bioinform.*, vol. 24, no. 1, Jan. 2023.

[134] 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.

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

[136] 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.

[137] 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.

[138] 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.

[139] 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.

[140] 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.

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

[142] 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.

[143] 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