Research on Explainability of Deep Neural Networks and Its Applications

Si Hua Robert Ip *

Shanghai Yue Kong Pao Senior Secondary School, Shanghai, China

* Corresponding Author Email: robertip0528@gmail.com

Abstract. In recent years, Artificial Intelligence (AI) models have surged in popularity and success, garnering much attention across many disciplines and applications. In particular, recent developments in Deep Neural Networks (DNN) have allowed them to match and even surpass human capabilities in many tasks. However, the complexity of deep neural networks causes them function like black boxes that are very difficult for humans to understand, thus making them untrustworthy and limiting their usefulness in high-risk applications. Explainable Artificial Intelligence aims to alleviate this issue by generating explanations for DNNs to allow humans to understand their reasoning and increase their transparency. This paper reviews research on neural network explainability, detailing its definition, necessity, classification, and evaluation. It categorizes current explanation methods for deep neural networks according to their number of inputs. Additionally, it introduces the principles and evaluation methods for common interpretability algorithms. Future research directions and applications of interpretable neural networks are summarized. The challenges facing interpretable neural networks are outlined, along with potential solutions to these challenges.

Keywords: Neural Networks; Interpretability; Explainable AI; Classification.

1. Introduction:

Deep Neural Networks (DNN) refers to the broad class of neural networks with multiple layers between the input layer and output layer, commonly trained with deep learning on large datasets to perform high-complexity tasks [1, 2]. Models such as Convolution Neural Networks (CNN) for image processing and Recurrent Neural Networks (RNN) for sequential data input have been shown to perform well at tasks such as Computer Vision, Speech Recognition, and Natural Language Processing (NLP). Recently, the Transformer architecture has further augmented the capacity and efficiency of DNNs in NLP with its multi-head attention mechanism, giving rise to powerful and commercially successful Large Language Models such as ChatGPT, Claude, Gemini, Llama and more.[3]

Despite their advantages, DNNs are still incomprehensible for human beings akin to a black box, even if the inner workings of the model are known. DNNs can have billions of parameters trained from machine learning, and it is infeasible for humans to interpret the meanings of them in an understandable way, especially since data is typically stored as abstract high dimensional vectors with no clear interpretation. As a consequence, it is very difficult for humans to trust the output, which poses many challenges in application.[4] The creator of the model cannot guarantee that the model has truly learnt meaningful knowledge from the data, and thus cannot guarantee its performance for all inputs. Users cannot trust the results of the model without knowing its reasoning, which especially limits its use cases in high-risk applications such as medical diagnosis and autonomous driving. Ethical concerns such as privacy, bias and discrimination also cannot be easily detected. Therefore, it is necessary to be able to produce explanations for the results that are both accurate with respect to the model and interpretable to humans, a topic known as Explainable Artificial Intelligence (XAI). The EU has also defined a “Right to Explanation of Decision Making” in their AI Act that requires relevant firms to be able to provide an explanation for a model’s output upon user demand, making explainable AI a necessity in the future.[5]

The existing research on explainable artificial intelligence is highly varied in their approaches at explaining models and how the explanations are presented to human users. Several surveys have attempted to classify them according to various metrics such as the type of explanation. This paper will describe a different form of classification based on whether the explanation is generated for a single input or generalized for multiple ones. Another often overlooked aspect is in the real-world application of XAI, which this paper will also aim to cover and provide an overview for current applications and the practical challenges it faces.

2. Definitions:

The term "Explanation" is not well-defined in most research, and is largely based on the human intuition of researchers and/or test subjects. What constitutes a valid explanation likely differs based on the person receiving the explanation due to differences in knowledge, domain experience, reasoning ability, and context. Broadly speaking, an explanation is simply any piece of information that assists humans in understanding a model's decisions accurately.

"Interpretability" and "Explainability" are two closely related concepts that have sometimes been used interchangeably to refer to any method of making a model understandable to humans. Existing research has several varying definitions for each, but for the sake of this paper they will be defined as such:

Interpretability refers to the ability for a model to be understood inherently.

Explainability refers to the ability to provide an understandable explanation.

Some authors include the act of providing explanations in Interpretable AI [6], while others view Interpretable AI as a subset of Explainable AI. [4] The confusion likely comes from the similar meanings of the two words, and the ambiguity as to what is considered interpretable and what constitutes an explanation. It is also worth noting that the term "Interpretable Machine Learning" (IML) precedes "Explainable Artificial Intelligence" (XAI), the latter of which remains more popular.

Other terms such as Trustworthy AI and Responsible AI are also used to describe different methods of bridging the gap between model and user, and have separate metrics from Interpretability and Explainability.[4] These are more commonly used when discussing the social, ethical, or legal aspects of AI, and as such will be beyond the scope of this paper.

The following terms are also commonly used when describing explanation methods. Their descriptions are based on definitions given by [2] and [7]:

Intrinsic means that the inner workings of the model is transparent when attempting to understand its reasoning. Interpretability is required to be intrinsic to the model, meaning that humans can directly access the model and understand its inner workings without additional assistance. This definition greatly limits the types of interpretable models to simple ones such as linear models or decision trees. DNNs are far too complicated to be intrinsically interpretable, and as such requires explanations. Explanation methods are considered intrinsic if they directly access the inner workings of the model when producing an explanation.

Post-hoc means that the explanation is produced for a black-box model solely by observing its inputs and outputs. It is generated after one or more inputs are given to the model and their corresponding outputs are returned. As a result, post-hoc explanations are less accurate in representing the true reasoning of the model's decision than intrinsic explanations, though that doesn't necessarily diminish its usefulness to humans.

Local explanations are only accurate for a small neighborhood of inputs. This is common when dealing with large models with a wide range of inputs, where the logic used for one input may be very different from that of a different one. These are often more popular due to their high accuracy and lower cost.

Global explanations are accurate across all inputs, and can explain the workings of the model as a whole. These are used to identify important trends across all inputs such as the most significant input variables used to provide predictions, and gain insight as to the general workings of the model.

However, they may not be accurate for each specific input and overlooks outliers that do not follow the general trends.

Model-specific explanations can only be for certain types of models, such as methods that are only suitable for specific modalities, or those that utilize the attention mechanism.

Model-agnostic explanations are applicable for all models, and are not restricted by the modality of inputs or outputs. As such, they are post-hoc by nature.

Modality refers to the way in which something is experienced or produced. The most common modalities include text, audio, visual, and tabular, though many others are often used in their specific domains, such as different types of medical data. Most existing research on explainability focuses on tabular and visual data.

3. Overview of Relevant Studies

3.1. Brief Description of Representative Studies

This section will provide brief descriptions for some of the most representative and popular methods currently used in XAI.

Class Activation Mapping (CAM) is an intrinsic, local, model-specific method used for CNN-based models that produces a saliency map of the most important regions contributing to the model predicting a specific class [8]. CAM trains a linear model on the feature maps obtained from CNN to obtain the weights of each feature's importance. Gradient-based Class Activation Mapping (Grad-CAM) is a generalization of CAM that uses the gradients flowing into the last convolutional layer as weights in the linear model to produce the saliency maps. Guided Grad-CAM is a type of Grad-CAM where guided backpropagation is used instead of normal backpropagation to obtain only the positive gradients that indicate a positive contribution to the output.

Local Interpretable Model-Agnostic Explanations (LIME) [9] produces an interpretable linear model that approximates the DNN locally around a given input. LIME perturbs samples of the input and obtains the model's output for each, then trains a linear model using these samples to train a linear model that is assumed to be interpretable to humans.

Shapley Additive Explanations (SHAP) [10] is a post-hoc, model-agnostic, global method that uses Shapley values from cooperative game theory to determine the influence of each input feature towards the final output. To do so, the model is formulated as a game, with the input features being the players that "cooperate" to obtain the final output.

Counterfactual Explanations are explanations in the form of "if X was X*, then output would be Z instead of Y" [11]. Instead of explaining why the model predicted what it did, it instead attempts to explain why it didn't predict something else. This is done by finding the minimum edits to the original input that would cause the model to change its prediction, hence locating the most discriminative features. Various approaches differ in the exact formulation of this problem and the techniques used during optimization. Contrastive Explanation Method (CEM) [12] is a type of counterfactual explanation that solves for both "Pertinent Positives" that are the minimum features required for producing the output, and "Pertinent Negatives" that are the minimum features required to be absent to prevent the model from switching to a different prediction.

Anchors [13] are post-hoc, local, model-agnostic explanations that provide key if-then rules that determine the prediction of the model. If an anchor's condition is fulfilled, the prediction is "anchored" in place and there is a very high probability that its consequent will also be true. Anchors are strictly causal relationships between input features and output predictions.

3.2. Relevant Synthesis Studies

As the amount of novel research and methodologies increased, numerous studies have also attempted to summarize and build a theoretical foundation to this field. Many studies have focused on the taxonomy and classification of popular methods and providing a scientific evaluation of each. [3] summarized the aforementioned taxonomy in Section 2, and created a database of existing

methods categorized based on its taxonomy. The study also evaluated the most popular methods within each modality on multiple datasets. [14] focuses on the existing explanation methods for multimodal neural networks, where it classified explanation methods based on the both the number of modalities of model input and the number of modalities of model output. It compares unimodal explanations with their multimodal counterparts and discusses the challenges in applying existing unimodal methods into multimodal networks. [15] focuses on visualization techniques of CNNs, where visual explanations are generated for CNNs with image input. It summarizes existing explanation methods, evaluation methods, and discusses applications of visualizations in the real world. Instead of summarizing preexisting research, [6] proposes a purely theoretical framework of classifying evaluation methods depending on whether it's Application-grounded, Human-grounded, or Functionally-grounded.

4. Classification Methods of Novel Explanation Algorithms

There are many different methods of classifying algorithms based on the different components of the explanation method. [16] categorizes explanations by 4 factors: their Model Dependency (Agnostic or Specific), Scope (Local or Global), Stages (Pre-, Post-, or during modelling), and Working Principle (Visualization, Distillation, or Intrinsic).

Visualization-based explanations are used to visually represent the most relevant features in the form of a saliency map, and are commonly used due to their intuitive and easily understandable presentation. Distillation explanations aim at extracting key knowledge from the model into a separate, interpretable model. Intrinsic Explanations use easily interpretable parts of the model such as the Attention mechanism to generate explanations; note that this use of the word "intrinsic" is different from the previous definition where it was used in opposition to "post-hoc", which is a commonly used metric not mentioned in this paper.

Explanation methods are also often categorized based on the type of explanation they produce. [17] Saliency maps are produced by visualization methods and show the relevance of different features or regions in the input towards the final prediction for images. For tabular data, saliency is found by training a separate surrogate model that is inherently interpretable such as linear models use the weights in the surrogate to obtain saliency. They are popular as they are easily understandable, but cannot be used to extract knowledge from the model as they do not provide a clear logical relationship between input and output. As such, they are more suitable for error and bias detection such as in image classification, where humans already have a good ground truth understanding and can verify that through the saliency maps that the model has learnt the correct knowledge. To solve this, one approach is to obtain causal rules in the form of "if-then" statements; however, it is impractical to fully explain any prediction made by a large enough model due to the complexity involved, limiting its use cases. Another approach is to obtain counterfactual explanations, where instead of explaining why the output was X, it is more practical to explain why it was not Y, and find the key difference that changes the prediction. Some methods attempt to understand the model indirectly by providing prototypes, which are representative inputs similar to the explained one given to the human in hopes that the human brain can deduce logical patterns from it, similar to how children learn by looking at examples of something until they form the correct connection. Finally, natural language explanations are an enticing possibility, which has advantages in its user-friendliness and understandability [18] . However, given that advanced NLP models are also black boxes, the accuracy of such explanations itself cannot be fully guaranteed, resulting in the same problem.

When applying explanation methods in the real world, one important factor to consider is whether the explanation requires multiple input or is only dependent on a single input. Most current methods focus on explaining the logic behind a single given input, and are used to justify the prediction of the model. These are local in nature, meaning that they can only be used on a case-by-case basis, but they require much less computation power. For a more general understanding of the model, multiple inputs are required to extract patterns and features. Whereas single input explanations can only identify the

most significant parts of the input (e.g., pixels in an image), multi-input explanations can combine individual features into general, abstract attributes and concepts (e.g., the stripes of a zebra), allowing users to directly gain high-level insight. The following section will provide examples for both single input and multi-input methods.

5. Classification of Explanation Methods Based on Input

5.1. Explanation Methods based on Single Input

5.1.1 Class Activation Mapping

Class activation mapping (CAM) [8] is an intrinsic, model-specific method of generate class activation maps that highlight the most discriminative regions contributing to a particular class within convolutional neural networks. Numerous studies have shown that the convolutional layers of CNNs possess a significant ability to extract relevant features and localize objects, but this ability is often diminished when fully connected layers are used for classification. To preserve its localization capabilities, CAM directly alters the network by forgoing the fully connected layer entirely and replacing it with a Global Average Pooling layer. Since the last convolutional layer should produce the most accurate and significant feature maps, a class activation map can be created by identifying the most salient features and projecting it back onto the pixels of the input image. This is done by assigning weights to each feature map and training these weights as a linear model for a particular output class. The resulting class activation map therefore shows how important each pixel is towards a class. Experiments confirm that CAM can maintain network classification performance while achieving localization. However, the process of altering the original network and retraining a linear model for each class is undesirable. To solve this, Grad-CAM forgoes the linear model used to train the weights, and instead uses the gradients of the model as the weights to the feature maps.

5.1.2 Gradient and Backpropagation-based Methods

Gradient-based methods rely on the gradients of the model at some or all layers to generate saliency maps. In theory, gradient is directly correlation with saliency, such that the higher the gradient, the more important it is to the output, though in practice the results obtained can be very noisy. Gradients are calculated by applying backpropagation from the output layer to the input layer, where the gradients are then calculated for every pixel to produce a saliency map. [19] introduces the SmoothGrad method, which optimizes saliency maps by adding smooth gradients to reduce visual noise. Since noise in saliency maps are often caused by insignificant but rapid fluctuations of the partial derivatives, SmoothGrad adds random noise to the gradient and computes a local average, resulting the final map being sharper and less noisy. [20] introduces Integrated Gradients, which is the path integral of the gradients along a straightline from a predetermined baseline, such as a black image, and the input. Intuitively, it measures the degree to which each pixel alters the prediction of the model from the baseline to be that of the input. It satisfies several axioms defined in the literature such as Sensitivity, Implementation Invariance, Linearity, and Completeness. [21] discusses two key assumptions often relied upon in saliency map interpretations: the use of first-order approximations of the loss function and the evaluation of each feature's significance in isolation. By retaining the Hessian term in the Taylor expansion and calculating it in a closed form, this approach reduces noise in saliency maps and aligns them more closely with original image objects.

5.2. Explanation Metrics based on Multiple inputs

Multi-input explanation methods aim to provide a unified explanation for the output of similar input samples by extracting generalized features found within the samples and describing each feature's contribution to the network's output. A straightforward approach involves generating individual explanations for each sample using a single input method first, then using a different model to extract similarities and patterns. Literature [22] highlights how many current machine learning algorithms suffer from the "Clever Hans" phenomenon, where models learn spurious correlations

rather than true intelligence, such as from coincidental patterns in the dataset. To address this and detect for false correlations, it proposes the SpRAy (spectral relevance analysis) algorithm, where individual attributes are clustered and summarized. This algorithm first computes individual relevance maps between input samples and the target category, then performs eigenvalue-based cluster analysis to identify different prediction strategies and patterns.

Another approach is to structurally create multiple levels of explanations, similar to how humans think in terms of different levels of abstraction. MAME (model-agnostic multilevel explanations) [23] is a meta-methodology that constructs a multilevel explanation tree based on existing local explanations. The tree's bottom level corresponds to local explanations, the top level to global explanations, and the intermediate levels to explanations of clustered data sets. Local interpretability techniques like LIME are used as the base for the tree's leaves. These explanations are then aggregated and grouped together based on their similarity, with the assumption that similar inputs produce similar explanations. Each cluster then produces its own explanation according to the patterns in the cluster. This process is then recursively performed on the new explanations to produce higher level explanations, eventually producing one global explanation. This method is validated in different scenarios with and without expert knowledge, and outperformed other methods in metrics such as fidelity. To globally explain black-box models, literature [24] introduces GIRP (global interpretation via recursive partitioning), which uses binary trees to clarify implicit decision rules. GIRP first generates a contribution matrix of each input variable for each sample using a single-input explanation method. Then, the model optimizes for a variable v and a rule by calculating the average contribution of v to each subset's prediction and finding the maximum. This process is recursively applied to produce a binary tree that approximates the original model while being easily interpretable. This algorithm has been tested to be versatile across different tasks such as scene understanding, text categorization, and ICU mortality prediction.

5.3. Evaluation Metrics for Explainable Algorithms

Evaluation Metrics [25] attempt to measure the effectiveness or quality of explanations. Since the end goal of explanations is for human users to gain insight into the workings of the model, an objective measurement is very difficult when measuring the quality of the explanation, and for practical applications expert or user feedback is often times essential and most effective. Subjective measurements such as Trust, Persuasiveness, Satisfaction, and Usefulness depend on human experience and can only be measured scientifically with user testing, and delves into the field of user psychology [26], which is outside the scope of this paper. Instead, the following section will focus on objective metrics only, with no consideration of the "goodness" of the explanation.

Stability, or Robustness, is the quality of an explanation being consistent across similar inputs and resistant to noise [27] This is simply measured by perturbing an input and measuring the change in explanation. High stability or robustness means that similar inputs should produce similar explanations, so the change in explanation should be small. This is especially important for making the explanation method robust against malicious attacks, which may attempt to perturb inputs in a way that is unnoticeable to humans but changes the explanation greatly.

Fidelity [28, 29]is the degree to which an explanation correctly describes the model's true behaviour. High fidelity means that a human with the explanation can accurately predict the model's behaviour given a similar input. For saliency-based explanations, two methods used are Deletion and Insertion. For a given input and the corresponding prediction, Deletion deletes each feature of the input one by one and measures the change in output, whereas Insertion starts from a blank input and adds each feature one by one and measures the change in output. A high saliency feature should significantly push the output towards the correct prediction when inserted, and significantly push it away from the prediction when deleted, and vice versa for low saliency features. A subcategory of fidelity is monotonicity, which is the idea that when features of increasing importance are added, the result must also be increasing in performance, else it would mean that a higher saliency feature contributes less than a lower saliency one, indicating an error. Fidelity does not measure the factuality

of the explanation, only that the saliency values are accurate, and hence is only useful for verifying the validity of an explanation.

Ground truth information is factual information about the subject in the real world. When ground truth information is provided, objective measurements on the factuality of explanations can be obtained [12]. For visual explanations, one common metric is Intersection over Union (IoU). Both the ground truth information and the explanation are presented as reference masks or bounding boxes of the most significant section(s) of an image. IoU is simply the ratio between the area of intersection and the area of union, with high IoU indicating that the explanation is truthful. For textual explanations, existing metrics in NLP such as Bilingual Evaluation Understudy (BLEU) can be applied, which simply measures the similarity between 2 sentences, in this case the computer-generated explanation and the ground truth explanation.

On the application side, there are also practical considerations when applying an explanation method. Given future regulations requiring companies to be able to provide explanations for their models, it may be necessary to be able to quickly produce many local explanations of user inputs. In that case, computational cost and efficiency are two important metrics to consider.

6. Applications, Research Directions, and Challenges of Explainable Neural Networks

6.1. Applications

XAI is currently mostly researched for high-stakes applications with existing neural network tools, such as in healthcare, criminal justice systems, and autonomous driving. Due to its novelty and a lack of necessity, it has currently not seen much commercial success as a user product, rather mostly being used by engineers to assess and debug current neural networks [30].

In healthcare, neural networks are used in AI-based support systems for medical diagnosis. [31] describes the use of XAI methods for explaining AI diagnosis for images, MRI, and CT imaging. [32] describes the use of a multimodal AI to diagnose patients based on data including textual transcripts and descriptions, tabular data of laboratory measurements, visual data from imaging, and more complex medical data types that are longitudinal in nature. This presents a new challenge as traditional XAI methods struggle to deal with such a diverse set of inputs. The study proposes an LLM-based “XAI Orchestrator”, where individual XAI models for each type of data provides explanations to the Orchestrator, which aggregates all information and extracts relevant knowledge from a knowledge base, producing a textual explanation to the user.

Criminal justice systems, such as in the U.S., use neural networks as risk assessments to predict crime locations and potential reoffenders. However, given the history of crime in the U.S., models tend to become biased towards certain races and ethnicities, as crime had statistically been more common among those communities. For example, black defendants were assessed to be as high as double the risk compared to white defendants, and also had much higher false positive rates. Literature [33] describes how explanation methods can detect and correct biases to ensure fairness in the justice system. It proposes a new prototype-based approach that extracts local data around each prototype and produces causal if-then rules based on the similarity of an input to each prototype.

Autonomous driving systems operate in unpredictable and high-risk environments, while requiring very complex algorithms from objecting detection to decision making. Applying interpretability algorithms to visual understanding [34] and situational awareness [35] in autonomous driving enhances model transparency. Addressing anomalies on the road is vital, and literature [36] proposes a self-evolutionary model that actively learns from current situations and provides explanations, improving autonomous driving safety.

6.2. Challenges and Future Research Directions

Despite significant advancements, the following challenges are still pressing issues that remain largely unaddressed, and will hinder the future development of XAI:

Lack of formal terminology. As mentioned previously, there is uncertainty towards most of the field's terminology, with each study having their own definitions towards even the most fundamental concepts such as "Explainability" and "Interpretability". What exactly counts as an "explanation" is also vague and elusive; [37] provides 5 separate definitions from different philosophical theories, and examines how the ideas are implicit in current research. A set of formal, agreed-upon definitions of the key terms are essential for enabling further discussion and research.

Similarly, there is a lack of well-defined, widely accepted metrics for evaluating explanations, especially in regard to the quality of explanations, which is partly due to the lack of formal definitions on what constitutes an effective explanation. This hinders the ability for comparison across different methods, especially with those with different approaches.

Research in causal relationships. Many current explanations are lacking in the ability to provide causal logic, which greatly diminishes its usefulness as a method of understanding models. Causal relationships are integral to human's understanding of the world, but current models fail to provide strict causal rules; for example, saliency maps can only show what the most important features are, but not why they are important. [38] argues that even counterfactual explanations are not grounded in causality, and thus often produce spurious connections instead of true, accurate logic.

Lack of ground truth databases. Ground truth information is essential in verifying the correctness of explanations, but there are very few existing ground truth knowledge bases to be used as reference [39]. Ideally, datasets should be accompanied with ground truth explanations for each sample explaining the correct prediction and reasoning behind it. This would require large amounts of manual annotations done by experts, which is the main reason behind its absence, but will be able to enable much more accurate explanation and evaluation methods.

Increasing model complexity. State-of-the-art neural networks are increasingly complex with billions of parameters across many hidden layers, which is necessary for high performances. However, there exists a tradeoff between model complexity and explainability, as the more complex a model is, the harder and more resource-intensive it is to find a simple explanation of its workings. [40] points out how modern LLMs are especially problematic, both in their opacity and in their immense size, which makes calculating explanations very costly and impractical for multiple inputs. More efficient algorithms are therefore necessary to be able to properly explain LLMs in the future.

References

[1] Joshi, Gargi, et al. "A Review on Explainability in Multimodal Deep Neural Nets." IEEE Access, vol. 9, 2021, pp. 59800–21. [doi: 10.1109/ACCESS.2021.3070212.]

[2] Bodria, Francesco, et al. "Benchmarking and Survey of Explanation Methods for Black Box Models." Data Mining and Knowledge Discovery, vol. 37, no. 5, Sept. 2023, pp. 1719–78. [doi: 0.1007/s10618-023-00933-9.]

[3] Rangapur, Aryan, and Aman Rangapur. "The Battle of LLMs: A Comparative Study in Conversational QA Tasks". [arXiv:2405.18344]

[4] Kong XW, Tang XZ, Wang ZM. "A survey of explainable artificial intelligence decision." Systems Engineering —Theory & Practice, 2021, 41(2): 524–536 [doi: 10.12011/SETP2020-1536]

[5] Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). "A Review of Trustworthy and Explainable Artificial Intelligence (XAI)." IEEE [doi: 10.1109/ACCESS.2023.3294569]

[6] Doshi-Velez, Finale, and Been Kim. "Towards A Rigorous Science of Interpretable Machine Learning" 2 Mar. 2017. [arXiv: 1702.08608].

[7] Hassija, Vikas, et al. "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence." Cognitive Computation, vol. 16, no. 1, Jan. 2024, pp. 45–74. [doi: 10.1007/s12559-023-10179-8.]

[8] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization." International Journal of Computer Vision, 128(2), 336–359. [doi: 10.1007/s11263-019-01228-7]

[9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier [arXiv: 1602.04938]

[10] Lundberg, Scott, and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." 24 Nov. 2017. [arXiv: 1705.07874].

[11] Goyal, Yash, et al. "Counterfactual Visual Explanations." 11 June 2019. [arXiv: 1904.07451].

[12] Dhurandhar, Amit, et al. "Explanations Based on the Missing: Towards Contrastive Explanations with Pertinent Negatives." 29 Oct. 2018. [arXiv: 1802.07623].

[13] Ribeiro, Marco Tulio, et al. "Anchors: High-Precision Model-Agnostic Explanations." Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, Apr. 2018.: [doi: 10.1609/aaai. v32i1.11491.]

[14] Rodis, Nikolaos, et al. "Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions." 30 June 2024. [arXiv: 2306.05731].

[15] Si NW, Zhang WL, Qu D, Luo XY, Chang HY, Niu T. "Representation visualization of convolutional neural networks: A survey." Acta Automatica Sinica, 2022, 48(8): 1890–1920 (in Chinese with English abstract). [doi: 10.16383/j.aas.c200554]

[16] Joshi, Gargi, et al. "A Review on Explainability in Multimodal Deep Neural Nets." IEEE Access, vol. 9, 2021, pp. 59800–21. [doi: 10.1109/ACCESS.2021.3070212.]

[17] Adadi, Amina & Berrada, Mohammed. (2018). "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". [doi:10.1109/ACCESS.2018.2870052.]

[18] Park, Dong Huk, et al. "Multimodal Explanations: Justifying Decisions and Pointing to the Evidence". [arXiv :1802.08129]

[19] Smilkov, Daniel, et al. SmoothGrad: "Removing Noise by Adding Noise." 12 June 2017. [arXiv: 1706.03825].

[20] Sundararajan, Mukund, et al. Axiomatic Attribution for Deep Networks. 12 June 2017. [arXiv: 1703.01365].

[21] Singla S, Wallace E, Feng S, Feizi S. "Understanding impacts of high-order loss approximations and features in deep learning interpretation." In: Proc. of the 36th Int'l Conf. on Machine Learning. Long Beach: PMLR, 2019. 5848–5856.

[22] Lapuschkin, Sebastian, et al. "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn." Nature Communications, vol. 10, no. 1, Mar. 2019, p. 1096. [doi: 10.1038/s41467-019-08987-4.]

[23] Ramamurthy, Karthikeyan Natesan, et al. "Model Agnostic Multilevel Explanations." 12 Mar. 2020. [arXiv: 2003.06005].

[24] Yang, Chengliang, et al. "Global Model Interpretation via Recursive Partitioning." 23 May 2018. [arXiv: 1802.04253].

[25] Coroama, Loredana, and Adrian Groza. "Evaluation Metrics in Explainable Artificial Intelligence (XAI)." Advanced Research in Technologies, Information, Innovation and Sustainability, edited by Teresa Guarda et al., vol. 1675, Springer Nature Switzerland, 2022, pp. 401–13. [doi: 10.1007/978-3-031-20319-0_30.]

[26] Hoffman, Robert R., et al. "Metrics for Explainable AI: Challenges and Prospects". [arXiv:1806.08049]

[27] Alvarez-Melis, David, and Tommi S. Jaakkola. "On the Robustness of Interpretability Methods". [arXiv:1806.08049]

[28] Miró-Nicolau, Miquel, et al. "A Comprehensive Study on Fidelity Metrics for XAI". [arXiv:2401.10640]

[29] Phillips, P. Jonathon, et al. "Four Principles of Explainable Artificial Intelligence." NIST IR 8312, National Institute of Standards and Technology (U.S.), 29 Sept. 2021, p. NIST IR 8312. [doi: 10.6028/NIST.IR.8312.]

[30] Bhatt, Umang, et al. "Explainable Machine Learning in Deployment". [arXiv:1909.06342]

[31] Singh, Amitojdeep, et al. Explainable Deep Learning Models in Medical Image Analysis. 28 May 2020. [doi: 10.48550/arXiv.2005.13799.]

[32] Pahud de Mortanges, Aurélie, et al. “Orchestrating Explainable Artificial Intelligence for Multimodal and Longitudinal Data in Medical Imaging.” Npj Digital Medicine, vol. 7, no. 1, July 2024, pp. 1–10. www.nature.com, [doi: 10.1038/s41746-024-01190-w.]

[33] Soares E, Angelov P. “Fair-by-design explainable models for prediction of recidivism.” [arXiv: 1910.02043, 2019]

[34] Soares E, Angelov P, Costa B, Castro M. “Actively semi-supervised deep rule-based classifier applied to adverse driving scenarios”. In: Proc. of the 2019 Int’l Joint Conf. on Neural Networks. Budapest: IEEE, 2019. 1–8. [doi: 10.1109/IJCNN.2019.8851842]

[35] Soares E, AngelovPP, Costa B, Castro MPG,Nageshrao S, Filev D. “Explaining deep learning models through rule-based approximation and visualization.” IEEE Trans. on Fuzzy Systems, 2021, 29(8): 2399–2407. [doi: 10.1109/TFUZZ.2020.2999776]

[36] Gao JY, Wang XT, Wang YS, Xie X. “Explainable recommendation through attentive multi-view learning.” In: Proc. of the 33rd AAAIConf. on Artificial Intelligence and 31st Innovative Applications of Artificial Intelligence Conf. and 9th AAAI Symp. on Educational Advances in Artificial Intelligence. Honolulu: AAAI Press, 2019. 445. [doi: 10.1609/aaai.v33i01.33013622]

[37] Sovrano, Francesco, and Fabio Vitali. “An Objective Metric for Explainable AI: How and Why to Estimate the Degree of Explainability.” Knowledge-Based Systems, vol. 278, Oct. 2023, p. 110866. arXiv.org, [doi: 10.1016/j.knosys.2023.110866]

[38] Chou, Y.-L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2022). “Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms and applications”. Information Fusion, 81, 59–83.

[39] Miró-Nicolau, M., et al. “Assessing Fidelity in XAI Post-hoc Techniques: A Comparative Study with Ground Truth Explanations Datasets”. [arXiv:2311.01961]

[40] Singh, Chandan, et al. “Rethinking Interpretability in the Era of Large Language Models”. [arXiv:2402.01761]