Explainable Artificial Intelligence (XAI): Motivation, Terminology, and Taxonomy

Aviv Notovich, Hila Chalutz-Ben Gal, and Irad Ben-Gal

1 Motivation

Deep learning algorithms and deep neural networks (DNNs) have become extremely popular due to their high-performance accuracy in complex fields, such as image and text classification, speech understanding, document segmentation, credit scoring, and facial recognition. As a result of the highly nonlinear structure of deep learning algorithms, these networks are hard to interpret; thus, it is not clear how the models reach their conclusions and therefore, they are often considered black-box models. The poor transparency of these models is a major drawback despite their effectiveness. In addition, recent regulations such as the General Data Protection Regulation (GDPR), require that, in many cases, an explanation will be provided whenever the learning model may affect a person's life. For example, in autonomous vehicle applications, methods for visualizing, explaining, and interpreting deep learning models that analyze driver behavior and the road environment have become standard. Explainable artificial intelligence (XAI) or interpretable machine learning (IML) programs aim to enable a suite of methods and techniques that produce more explainable models while maintaining a high level of output accuracy [1–4]. These programs enable human users to better understand, trust, and manage the emerging generation of artificially intelligent systems [4].

Many people do not feel comfortable when blindly agreeing with an AI system's decisions in various situations, without some understanding of the decision-making process used by such a system. To achieve trust in AI systems, detailed "explana-

A. Notovich · I. Ben-Gal (✉)
Department of Industrial Engineering, Tel Aviv University, Tel-Aviv, Israel
e-mail: bengal@tauex.tau.ac.il

H. Chalutz-Ben Gal
School of Industrial Engineering and Management, Afeka Tel Aviv Academic College of Engineering, Tel Aviv, Israel

L. Rokach et al. (eds.), *Machine Learning for Data Science Handbook*,
https://doi.org/10.1007/978-3-031-24628-9_41

tions" of AI system decisions seem necessary. Such explanations provide insights into and interpretability of the rationale of the applied AI algorithms and help users trust the system conclusions. As ML and AI modeling are increasingly involved in critical areas such as transportation, retail, insurance, medicine, criminal justice, and financial markets, it seems vital that these models become more easily understood [9].

The XAI-related concepts of explainability, interpretability, and accuracy are presented next, followed by segmentation of XAI methods.

2 Explainability, Interpretability, and Related XAI Terms

The definitions of both AI explainability and AI interpretability have multiple meanings and sometimes there is little to no consensus in the research community regarding these terms [1]. There are a few conflicting definitions that differ from each other in terms of theme and community. In particular, various AI-related communities approach the concept of explainability from different angles. The term explainable AI (XAI) has a double meaning itself. Sometimes it is used to represent methods that help explore the mechanisms of the AI methods or the AI systems themselves; for example, a researcher may seek an interpretation of how these methods or systems work or which features are important when making predictions. In other cases, the term XAI is related to explanations about particular inputs, outputs and examples, such as understanding how a record in a dataset was mapped to a specific segment or recommendation.

Lipton [9] addresses this ambiguity and claims that many XAI papers provide diverse and sometimes non-overlapping motivations for interpretability and offer myriad notions of what makes render models interpretable. Despite such ambiguity, many papers proclaim interpretability axiomatically, absent further explanation.

Explainability and interpretability are closely related concepts in the literature. Sometimes, the term "explainability" refers to "why" a recommendation has been made, while the term "interpretability" refers to "how" that recommendation was obtained [2]. Accordingly, it has been claimed that interpretability is one of the approaches that achieves explainability [3]. Explainable AI (XAI) aims to develop tools that are able to explain AI model decisions to inexpert users. To do so, the model might be either interpretable or non-interpretable. Interpretable models try to develop models whose decision mechanism is locally or globally transparent. Therefore, the model outputs are usually naturally explainable.

Other approaches claim that "Explainability" and "Interpretability" are two related, yet distinct, concepts when referring to AI systems.

"AI Explainability" refers to the ability of an AI or ML model to provide understandable and clear explanations for its predictions or decisions. An explainable model should be able to articulate the reasons behind the model outputs in an easy and comprehended way by human users. For example, explainability is particularly important in domains where the impact of AI decisions can have legal, ethical, or societal consequences (e.g., healthcare, finance, and autonomous vehicles). An explainable model contributes to trust the model's decision-making process.

"AI Interpretability," on the other hand, refers to the ability of a model to be understood by humans in terms of its internal robustness or how it arrives at its outputs. An interpretable model is one that can be explained in terms of its feature importance, decision rules, or other transparent representations, which allow humans to understand how the model arrives at its predictions. Interpretability is often used interchangeably with explainability, but it can also refer specifically to the technical characteristics of a model that make it transparent and understandable.

Even though interpretability and explainability have been used interchangeably, Došilović et al. [4] claim it is important to distinguish between them. As such, explainable models are interpretable by default, but the reverse is not always true. However, interpretability is not the only way to achieve explainability. There are models that reveal their internal decision mechanisms for explanation purposes and use complex explanation techniques, such as neural attention mechanisms [3].

According to Došilović et al. [4], interpretability alone is insufficient. To increase human trust in black-box methods, it is necessary to develop explainability models that summarize the reasons for the model output. The authors assume that, while both mechanisms are important, interpretability is a substantial first step that provides the capacity to defend model actions and recommendations, provide relevant responses to questions, and be audited.

Interpretable models encompass much of the present work in explainable AI [1]. The main reason is the increased usage of deep neural networks that are so hard to interpret. However, it is still challenging to formulate a line of reasoning that explains a model's decision-making process to the user while relying on human-understandable features of the input data. Nonetheless, reasoning is a critical step when formulating an explanation about why or how an AI-based recommendation has been made.

To summarize the above discussion, despite the inherent inconsistency that one can find in the literature, the following list presents some of the common terms and their popular explanation in the XAI community. The list is mainly based on [2–4] that provide an excellent overview on the topic.

- **Interpretability –** users should be able to understand and reason about the model output.
- **Model Transparency –** defined in terms of *simulatability, decomposability,* and *algorithmic transparency.*
- **Simulatability –** whether a human can use the input data together with the model to reproduce every calculation necessary to make the prediction.
- **Decomposability –** whether there is an intuitive explanation of all the model parameters.
- **Algorithmic Transparency –** an ability to explain how the learning algorithm works.
- **Model Functionality –** defined in terms of *textual description, visualization,* and *local explanation.*

- **Textual Description –** a semantically meaningful description of the model output.
- **Visualization –** a method for explaining a model through visualization of its output and its parameters.
- **Local Explanation –** rather than explaining the mapping of an entire model, local changes are introduced using specific input vectors for a given output class. Explanation is provided on specific use cases or instances.
- **Global Interpretability –** understanding the entire ML model behavior, holistic reasoning that leads to all different possible outcomes.
- **Local Interpretability –** understanding a single model prediction.
- **Activation Maximization –** generation of an input image that maximizes the filter output activations.
- **Anchor –** rule that sufficiently "anchors" the prediction locally such that changes to the rest of the instance's feature values do not matter.
- **Surrogate Model –** a simple model on top of (or besides) a complex model, trained based on the same input and the same predictions of the original complex model in order to mimic a better explanation and interpretation.
- **Partial Dependence Plot (PDP) –** a graphical representation that helps visualize the average partial relationship between one or more input variables and the predictions of a complex model.
- **Individual Conditional Expectation (ICE) –** a graphical representation that reveals interactions and individual differences by separating the PDP output.
- **Knowledge Extraction –** the task of extracting explanations/knowledge from the complex model during training and encoding that knowledge as an internal representation of a complex model.
- **Influence Methods –** several techniques that carefully modify the inputs and measure how much the prediction changed according to each modification.
- **Example-Based Explanation –** selection of specific data points to explain the behavior of machine learning models.

3 Accuracy and Explainability

A conventional statement is that there is an inherent trade-off between model explainability and model effectiveness, thus stating that one can either achieve high explainability with simpler models or high accuracy with more complex models, which are generally harder to interpret [3]. Figure 1 presents a possible schematic view of the trade-off between the model explainability and the model effectiveness. Similar graphs can be found in many papers, with the same message.

This belief raises a common dilemma among practitioners regarding whether to choose an understandable/explainable simple algorithm, while sacrificing prediction accuracy or to choose an accurate latent factorization modeling approach, while sacrificing explainability [5]. However, there is also a belief that these two goals

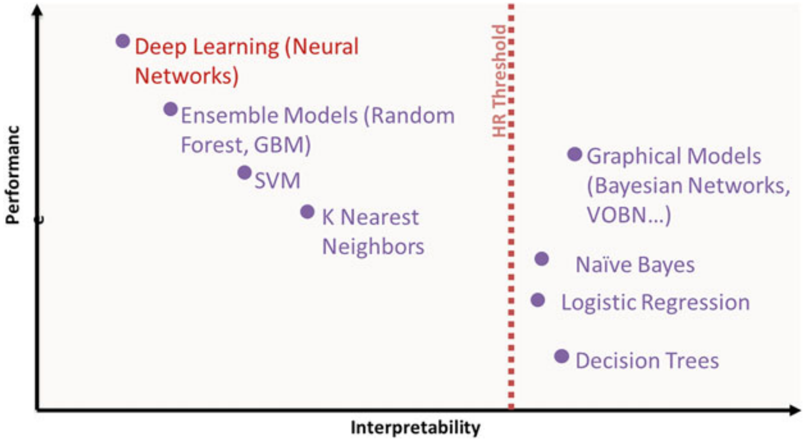

Fig. 1 The inherent trade-off between model performance to model interpretability

do not necessarily contradict each other [6], which claims that this assumption is primarily relevant for cases related to structured data with meaningful features.

Extensive research has focused on state-of-the-art techniques, such as deep learning approaches, which emphasize a model design that is both effective and explainable. Developing explainable deep models is thus an attractive direction in the broader AI community, leading to progress in essential explainable machine learning problems [3].

4 Segmentation of XAI Approaches

There are several ways to classify and segment the different XAI approaches [1, 2, 4]. Adadi and Berrada [3] propose a categorization for XAI methods that considers the model's complexity of interpretability, scope of interpretability, and level of dependency. In the next sections, we follow earlier surveys by Chakraborty [2], as well as Adadi and Berrada [3].

4.1 Complexity-Related Methods

Many works in the literature assume that model complexity is directly related to interpretability. Thus, simpler models are easier to interpret. Accordingly, to better interpret complex models, there is a need to introduce a simpler surrogate model or an algorithm for interpretability. Several works following this direction are described in this section.

Xu et al. [8] consider the task of automatically generating image captions as a goal that is central to scene understanding. The authors introduce an attention-based image caption model that automatically learns how to describe image content. They train the attention model using standard backpropagation techniques over deep neural networks and by maximizing a variational lower bound. The proposed model gains insight and interpretation by visualizing "where" and "what" the attention is focused on. Relying on visualization and benchmark datasets, they demonstrate how their model is able to interpret the images.

Caruana et al. [7] deal with pneumonia risk prediction by applying generalized additive models with pairwise interactions (GA^{2}Ms). The proposed model achieves state-of-the-art accuracy and is able to uncover surprising patterns in the data that previously challenged researchers and prevented the implementation of complex machine learning models in this domain. The model is used to identify and remove such patterns to obtain a better performance.

Letham et al. [6] propose a method based on decision trees called Bayesian Rule Lists (BRL), which produces a predictive model that is not only accurate but also interpretable to human experts. This model generates conditional "if/then" statements (e.g., "if high blood pressure, then stroke") that discretize a high-dimensional, multivariate feature space into a series of simple, readily interpretable decision statements. Such an outcome is highly interpretable and provides concise and convincing capabilities that are able to gain the trust of domain experts.

Lipton [9] proposes following a post hoc explanation approach with two stages. This approach first allows complex, uninterpretable black-box models to generate high-performance outputs, and then it applies a separate set of techniques to obtain explainability and interoperability over the outputs. Such an approach views the interpretability task as a reverse engineering process that provides the required explanations without altering or even knowing the inner works of the original black-box model.

4.2 Global and Local Interpretability Approaches

There are two primary approaches when seeking explainability and interpretability for AI and ML models. The first is the global interpretability approach, which aims to provide a systematic view and general understanding of the AI system in use. Thus, the global interpretability approach seeks a complete view of the decisions and operations of the entire AI model. For example, this approach focuses on explaining the overall model analysis using a set of rules and measures that determine the global feature importance and explain the model outcomes. Such explainability could be used for example by technical experts to obtain a better modeling decision.

The second is the local interpretability approach, which is focused on approximating and explaining individual predictions and case-by-case outcomes. Thus, unlike global interpretability, it does not seek to explain the whole model but rather the specific outcomes of the AI system under different feature values and conditions.

For example, local interpretability can be used to explain and justify an AI system recommendation that a specific client is not entitled to a bank loan due to his income level and previous loans or other personal financial conditions. As such, global interpretability is often better for non-technical users.

Note that some studies aim to combine global and local interpretability; examples of this approach include Guidotti et al. [21] and Linsley et al. [22].

Global Interpretability

As indicated, the goal of the global XAI approach is to understand the entire logic of a model and the entire pattern of reasoning that leads to different possible outcomes. This approach is most relevant in situations that require a high level of accountability and justification, such as AI applications in medical domains [1]. In these cases, a global effect estimate is often more helpful than many separate explanations for different possible predictions. Some examples that imply such an approach are as follows.

Nguyen et al. [12] aim to study what each of a DNN's neurons is learning to detect. They use activation maximization (AM), which synthesizes an input (e.g., an image) that highly activates a neuron. The proposed method generates synthetic images and reveals the features learned by each neuron in an interpretable way.

Valenzuela-Escárcega et al. [11] propose a supervised approach for information extraction, which combines bootstrapping with representation learning. The proposed algorithm iteratively learns custom embeddings for multi-word entities and their matched patterns from example entities for each classification category. This approach outputs a globally interpretable model consisting of a decision list that acts as an interpretation of the model.

Yang et al. [10] propose a method that interprets black-box machine learning models globally using a binary interpretation tree. The interpretation tree explicitly represents the most important decision rules that are implicitly contained in the black-box machine learning models. The proposed learning algorithm partitions the input variable space by maximizing the difference between the average contributions of the split variable over the divided spaces. This method results in a contribution matrix that consists of the contributions of input variables to the predicted scores for each single prediction. The authors demonstrate the effectiveness of their method for diagnosing machine learning models over multiple tasks as well as for analyzing the models in terms of human understanding.

Local Interpretability

The goal of local interpretability is to explain the reasons for a specific decision or single prediction that the ML model has made. Here, we discuss research works focused on this type of interpretability.

Ribeiro et al. [13] proposed a novel technique to explain the predictions of a classifier in an interpretable and faithful manner, by *locally* learning an interpretable model based on individual predictions. They called it the *local interpretable model-agnostic explanation* (LIME). The method, which is formulated as a submodular optimization problem, approximates a black-box model locally in the neighborhood of any prediction.

Ribeiro et al. [14] extend LIME using decision rules called “anchors”. An anchor explanation is a rule that sufficiently “anchors” the prediction locally, such that changes to the rest of an instance’s feature values do not affect the AI system recommendation.

A similar approach was used in a series of studies [15–19] that analyzed image classification by a family of ML models. In particular, the analyses identified image regions (pixels) that were found to be particularly influential on the final classification. Several names were given to this approach, including *sensitivity maps*, *saliency maps*, or *pixel attribution maps*. These techniques assign an “importance” score to individual pixels, which is meant to reflect their influence on the final classification of the image. A similar yet opposing concept is applied in *adversarial learning*, which aims to find and modify these specific pixels as a means to distort and change the correct classification [40].

Lundberg and Lee [20] present a unified framework for interpreting predictions, named SHAP (SHapley Additive exPlanations). SHAP assigns each particular prediction’s features an importance value. Its novel components include (i) the identification of a new class of additive feature importance measures and (ii) theoretical results showing that there is a unique solution in this class with a set of desirable properties. Additionally, the authors show that by using different kernels, SHAP can be model agnostic.

According to surveys by Chakraborty et al. [2] and by Adadi and Berrada [3], local explanations are the most commonly used explanation methods in XAI and are particularly applied to DNN models.

4.3 Model-Related Methods

Another popular way to classify model interpretability techniques is according to whether they are model agnostic or model-specific; model-agnostic methods can be applied to any model type, while model-specific methods work only for specific models.

Model-Specific Interpretability

As model-specific techniques are limited to a particular model; according to [2, 3], they are less popular than model-agnostic interpretability methods, which often generate more interest.

Model-Agnostic Interpretability

According to Mary [4], a specific class of model-agnostic methods is related to those that can be applied primarily to black box models’ inputs and outputs. The usability and popularity of these methods can be found by examining a variety of use cases [5]. This class of methods addresses prediction tasks and explanation tasks separately. Model-agnostic interpretations are usually post hoc, i.e., they are generally applied to interpret DNNs and could be either local or global interpretable models [3]. Herein, we present an overview of the studies focused on model-agnostic interpretability, grouped by the applied techniques. In particular, one can find four primary technique types: *visualization, knowledge extraction*, *influence methods,* and *example-based explanation* [3].

4.3.1 Visualization

One way to illustrate and better understand an ML model output, especially a DNN, is to represent it visually; for example, researchers have previously explored hidden patterns within a segment of the neural network (including a single neural unit). Many visualization techniques are applied to supervised learning models in which the active neurons and pixels can be highlighted per labeled class. The literature contains three primary types of explainability techniques that are related to visualization: *Surrogate models*, *Partial Dependence Plots* (PDP), and *Individual Conditional Expectation* (ICE) [3].

Surrogate Models

Surrogate modeling refers to building a simple model (e.g., a linear model or decision tree) to approximate a more complex model (e.g., a DNN) to help explain how the complex model reaches its decisions. To build a surrogate model, one should often train the simpler model based on the inputs and the outputs of the more complex original model. In many cases, the simpler model's output can be visualized to further highlight the important features on the model output. This technique is sometimes useful; however, there is no theoretical guarantee that this technique will produce a clean and effective explanation for the complex model.

LIME [13] is a popular method for constructing local surrogate models around subsets of observations. Bastani et al. [23] built such a surrogate model approach by extracting a decision tree that represents a complex model's behavior. Thiagarajan et al. [24] proposed an approach for building the "TreeView" representation using a surrogate model that performs hierarchical partitioning of the feature space. This surrogate model reveals the iterative rejection of unlikely class labels until the correct association is predicted.

Partial Dependence Plot (PDP)

The partial dependence plot is another graphical representation that helps visualize the average partial relationship between one or more input variables and a complex model's predictions. PDP has been used in several studies to understand the relationship between predictors and inputs under several conditions (e.g., [25–27]).

Individual Conditional Expectation (ICE)

Individual conditional expectation (ICE) can be considered an extension of PDP. ICE plots reveal interactions and individual differences by separating the PDP output. ICE has been used in several studies (e.g., [28, 29]), in which the advantage of ICE over PDP has been demonstrated and analyzed.

4.3.2 Knowledge Extraction

Knowledge extraction (KE) refers to the task of extracting explanations and knowledge from a complex model during the training phase and encoding it as an internal representation of a complex model. In the literature, two primary KE techniques include *rule extraction and model distillation* [3].

Rule Extraction

Rule extraction (RE) aims to find rules that provide approximation of the decision-making process for a more complex model. In a sense, it is similar to the association rules that were used in data-mining tasks to extract simple rules from ML classification models.

Using RE, one can obtain a better description of the knowledge learned by the complex model during training. Several studies have implemented rule extraction (e.g., [30, 31]).

Model Distillation

Model distillation (MD) is based on model compression techniques. MD was originally proposed to reduce the computational cost of a model at runtime but was later targeted at interpretability. Distillation is a model compression that transfers information from deep networks to shallow networks in the form of “teacher to student” [32]. Several studies have implemented model distillation (e.g., [33, 34]).

4.3.3 Influence Methods

Influence methods refer to several techniques that systematically modify a model’s inputs and then measure how much the prediction changed according to each modification. In this way, a relevance score for each feature is computed. According to [3], the literature describes three alternative methods for obtaining the input variable’s relevance: *sensitivity analysis*, *layer-wise relevance propagation,* and *feature importance*. These approaches are discussed next.

Sensitivity Analysis

Zhang and Wallace [35] introduce a sensitivity measure that determines how a complex model's output is influenced by its input and/or weight perturbations. In particular, they conducted a sensitivity analysis examining one-layer convolutional neural networks (CNNs) to explore the effect of architecture components on model performance; their aim was to distinguish between important and comparatively inconsequential design decisions for sentence classification.

Sensitivity analysis (SA) is widely used to verify whether model outputs remain stable when the data are changing and to support robustness verification in general. Cortez and Embrechts [36] proposed a global SA (GSA) method, which extends the applicability of previous SA methods and several visualization techniques when assessing input relevance and effects on the model's responses. The authors demonstrate the GSA method's capabilities by conducting several experiments using an NN ensemble and SVM model and including both synthetic and real-world datasets. It is worth mentioning, however, that this approach produces an explanation only over the variation of the function values but not the function itself.

Layer-Wise Relevance Propagation (LRP)

Bach et al. [37] proposed a pixel-wise decomposition for nonlinear classifiers. This technique provides visualization of the contributions of single pixels to the predictions of kernel-based classifiers, which can be visualized using heat maps. The proposed technique focuses the analysis on regions of potential interest while tracing backward from the prediction to the input layer. Unlike sensitivity analysis, this technique explains the predictions relative to the state of maximum uncertainty.

Feature Importance

Feature importance provides a score for each input feature that represents its contribution to the predictions of a complex ML model. Basically, this technique generates a permutation of the input features and measures the corresponding model error. Features with high importance increase the model error more significantly when permutated than a feature with low importance. Fisher et al. [38] proposed a technique called model class reliance (MCR), which sets the range of feature importance values across several models for a pre-specified class. Casalicchio et al. [29] used SHAP values to generate a feature importance score for every input feature.

4.3.4 Example-Based Explanation

Example-based explanations (EBEs) are techniques that select particular data points from the dataset to explain the behavior of the ML/AI model. In the reviewed literature, the two primary EBE techniques are *prototypes and criticisms* and *counterfactual explanations* [3].

Prototypes and Criticisms

To avoid overfitting the learning model, a strong representation for the data points must be selected. Kim et al. [39] claim that although example-based explanations are often used to interpret highly complex distributions, prototypes alone rarely sufficiently represent the essence of the model complexity. Motivated by the Bayesian model criticism framework, they develop the MMD-critic, which efficiently learns prototypes and criticism designed to aid human interpretability. The authors evaluate the prototypes selected by MMD-critic using a nearest prototype classifier, demonstrating competitive performance when compared to baselines.

Counterfactual Explanations

Counterfactual explanations attempt to find the boundary at which the learning model will change its decision or recommendation with minimum conditions. This outcome is achieved without the need to describe the algorithm's full logic. Yuan et al. [40] noted that ML models are vulnerable to well-designed input samples, called adversarial examples. Adversarial examples may be invisible to humans but can fool a complex ML model and alter their decision with minimal change to the input. The authors review recent findings related to adversarial examples for deep neural networks, summarize the methods that generate adversarial examples, and propose a taxonomy for these methods.

5 Final Remark

Adadi and Berrada [3] provide an excellent summary of different XAI methods. In a summary taxonomy table, they classify various XAI models and techniques by analyzing whether they are intrinsic/post hoc, global/local, and model specific/model agnostic.

As claimed by the authors, XAI is a vital interdisciplinary research direction and a major building block in the AI ecosystem. The potential impact of XAI can affect various new applications in areas such as transportation, healthcare, military, retail, legal, finance, and well-being. Yet, despite its importance, XAI research is still unstructured, and the human aspects in it can be further studied.

References

1. Erico Tjoa, Cuntai Guan, “A Survey on Explainable Artificial Intelligence (XAI): towards Medical XAI” 2019. [Online] https://arxiv.org/pdf/1907.07374.pdf
2. Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M. Rao, Troy D. Kelley, Dave Braines, Murat Sensoyk, Christopher J. Willis, Prudhvi Gurram IBM T. J. Watson Research Center, Crime and Security Research Institute, Cardiff University, UCLA, IBM UK, Army Research Lab, Adelphi, Ozyegin University, BAE Systems AI Labs University College London “Interpretability of Deep Learning Models: A Survey of Results” 2017. [Online] https://orca.cf.ac.uk/101500/1/Interpretability%20of%20Deep%20Learning%20Models%20-%20A%20Survey%20of%20Results.pdf
3. Amina Adadi, Mohammed Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)” 2018. [Online] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590
4. Filip Karlo Došilović, Mario Brcic, Nikica Hlupic “Explainable Artificial Intelligence: A Survey” 2018. [Online] https://www.researchgate.net/publication/325398586_Explainable_Artificial_Intelligence_A_Survey
5. Jakob M. Schoenborn, Klaus-Dieter Althof, “Recent Trends in XAI: A Broad Overview on current Approaches, Methodologies and Interactions”. 2019. [Online] http://gaia.fdi.ucm.es/events/xcbr/papers/XCBR-19_paper_1.pdf
6. B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,” Ann. Appl. Statist., vol. 9, no. 3, pp. 1350–1371, 2015. [Online] https://arxiv.org/pdf/1511.01644.pdf
7. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, “Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission,” in Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2015, pp. 1721–1730. [Online] http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf
8. K. Xu et al., “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn. (ICML), 2015, pp. 1–10. [Online] https://arxiv.org/pdf/1502.03044.pdf
9. Z. C. Lipton, “The mythos of model interpretability,” in Proc. ICML Workshop Hum. Interpretability Mach. Learn., 2016, pp. 96–100. [Online] https://arxiv.org/pdf/1606.03490.pdf
10. C. Yang, A. Rangarajan, and S. Ranka. (2018). “Global model interpretation via recursive partitioning.” [Online] https://arxiv.org/pdf/1802.04253.pdf
11. M. A. Valenzuela-Escárcega, A. Nagesh, and M. Surdeanu. (2018). “Lightly-supervised representation learning with global interpretability.” [Online] https://arxiv.org/pdf/1805.11545.pdf
12. A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2016, pp. 3387–3395. [Online] https://arxiv.org/pdf/1605.09304.pdf
13. M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’: Explaining the predictions of any classifier,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144. [Online] https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
14. M. T. Ribeiro, S. Singh, and C. Guestrin, “Anchors: High-precision model-agnostic explanations,” in Proc. AAAI Conf. Artif. Intell., 2018, pp. 1–9. [Online] https://homes.cs.washington.edu/~marcotcr/aaai18.pdf
15. K. Simonyan, A. Vedaldi, and A. Zisserman. (2013). “Deep inside convolutional networks: Visualising image classification models and saliency maps. [Online] https://arxiv.org/pdf/1312.6034.pdf

16. M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in Proc. Eur. Conf. Comput. Vis. Zurich, Switzerland: Springer, 2014, pp. 818–833. [Online] https://cs.nyu.edu/~fergus/papers/zeilerECCV2014.pdf
17. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and O. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., June 2016, pp. 2921–2929. [Online] https://arxiv.org/abs/1512.04150
18. M. Sundararajan, A. Taly, and Q. Yan. (2017). “Axiomatic attribution for deep networks.” [Online] https://arxiv.org/pdf/1703.01365.pdf
19. D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. (2017). “SmoothGrad: Removing noise by adding noise.” [Online] https://arxiv.org/pdf/1706.03825.pdf
20. S. M. Lundberg and S. I. Lee, “A unified approach to interpreting model predictions,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 4768–4777. [Online] https://arxiv.org/pdf/1705.07874.pdf
21. R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. (2018). “Local rule-based explanations of black box decision systems.” [Online] https://arxiv.org/pdf/1805.10820.pdf
22. D. Linsley, D. Scheibler, S. Eberhardt, and T. Serre. (2018). “Globaland-local attention networks for visual recognition.” [Online] https://arxiv.org/pdf/1805.08819.pdf
23. O. Bastani, C. Kim, and H. Bastani. (2017). “Interpretability via model extraction. [Online] https://arxiv.org/pdf/1706.09773.pdf
24. J. J. Thiagarajan, B. Kailkhura, P. Sattigeri, and K. N. Ramamurthy. (2016). “TreeView: Peeking into deep neural networks via feature-space partitioning.” [Online] https://arxiv.org/pdf/1611.07429.pdf
25. D. P. Green and H. L. Kern, “Modeling heterogeneous treatment effects in large-scale experiments using Bayesian additive regression trees,” in Proc. Annu. Summer Meeting Soc. Political Methodol., 2010, pp. 1–40. [Online] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.190.3826&rep=rep1&type=pdf
26. J. Elith, J. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” J. Animal Ecol., vol. 77, no. 4, pp. 802–813, 2008. [Online] https://besjournals.onlinelibrary.wiley.com/doi/full/10.1111/j.1365-2656.2008.01390.x
27. S. H. Welling, H. H. F. Refsgaard, P. B. Brockhoff, and L. H. Clemmensen. (2016). “Forest floor visualizations of random”. [Online] https://arxiv.org/pdf/1605.09196.pdf
28. A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” J. Comput. Graph. Statist., vol. 24, no. 1, pp. 44–65, 2015, [Online] https://www.tandfonline.com/doi/abs/10.1080/10618600.2014.907095
29. G. Casalicchio, C. Molnar, and B. Bischl. (2018). “Visualizing the feature importance for black box models.” [Online] https://arxiv.org/pdf/1804.06620.pdf
30. U. Johansson, R. König, and I. Niklasson, “The truth is in there—Rule extraction from opaque models using genetic programming,” in Proc. FLAIRS Conf., 2004, pp. 658–663. [Online] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.4124&rep=rep1&type=pdf
31. T. Hailesilassie. (2017). “Rule extraction algorithm for deep neural networks: A review.” [Online] https://arxiv.org/abs/1610.05267
32. P. Sadowski, J. Collado, D. Whiteson, and P. Baldi, “Deep learning, dark knowledge, and dark matter,” in Proc. NIPS Workshop High-Energy Phys. Mach. Learn. (PMLR), vol. 42, 2015, pp. 81–87. [Online] http://proceedings.mlr.press/v42/sado14.pdf
33. S. Tan, R. Caruana, G. Hooker, and Y. Lou. (2018). “Detecting bias in black-box models using transparent model distillation.” [Online] https://arxiv.org/abs/1710.06169
34. Z. Che, S. Purushotham, R. Khemani, and Y. Liu. (2015). “Distilling knowledge from deep networks with applications to healthcare domain.” [Online] https://arxiv.org/abs/1512.03542
35. Y. Zhang and B. Wallace (2016). “A sensitivity analysis of (and practitioners’ Guide to) convolutional neural networks for sentence classification. [Online] https://arxiv.org/abs/1510.03820

36. P. Cortez and M. J. Embrechts, “Using sensitivity analysis and visualization techniques to open black box data mining models,” Inf. Sci., vol. 225, pp. 1–17, Mar. 2013. [Online] https://core.ac.uk/download/pdf/55616214.pdf
37. 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, p. e0130140, 2015. [Online] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140
38. A. Fisher, C. Rudin, and F. Dominici. (2018). “Model class reliance: Variable importance measures for any machine learning model class, from the ‘rashomon’ perspective.” [Online] https://arxiv.org/abs/1801.01489
39. B. Kim, R. Khanna, and O. O. Koyejo, “Examples are not enough, learn to criticize! criticism for interpretability,” in Proc. 29th Conf. Neural Inf. Process. Syst. (NIPS), 2016, pp. 2280–2288 [Online] https://papers.nips.cc/paper/6300-examples-are-not-enough-learn-to-criticize-criticism-for-interpretability.pdf
40. X. Yuan, P. He, Q. Zhu, and X. Li. (2017). “Adversarial examples: Attacks and defenses for deep learning.” [Online] https://arxiv.org/abs/1712.07107