Chapter 3
An Overview of Explainable AI Methods, Forms and Frameworks

Dheeraj Kumar and Mayuri A. Mehta

3.1 Introduction

Artificial Intelligence (AI) has become an integral part of many applications in recent years. AI has reached the masses with the availability of intelligent software to automate systems. Machine learning models like a decision tree and Bayesian network are easily interpreted by the realization of the importance of each feature to the output. However, models developed using a deep neural network (Shi Zhang and Zhu 2018; Ras et al. 2018; Gilpin et al. 2018) do not allow understanding its internal mechanism. Due to the lack of transparency in the deep neural networks, it is hard to justify their predictions to the end-users. In addition, the stochastic nature of predictions made by them is another obstacle to understandability and transparency for humans. Thus, explaining the model developed using a deep neural network (often referred to as the black box model) is essential for users to accept AI-based solutions (Inam et al. 2021).

A black box model does not disclose its internal design and inference mechanism while making a prediction. Moreover, explanations are also necessary to evaluate the ethical and moral standards of a machine (Islam et al. 2021; Doran et al. 2018). Explainable Artificial Intelligence (XAI) aims to create a suite of techniques and frameworks that explain and interpret predictions made by black box models. The explainability of a black box model is the ability to explain its prediction in an understandable form for end-users. The goal of XAI is to communicate to end-user why a

D. Kumar (✉)
Department of Information Technology, Parul Institute of Engineering and Technology, Parul University, Vadodara, India
e-mail: dhirajsingh66@gmail.com

M. A. Mehta
Department of Computer Engineering, Sarvajanik College of Engineering and Technology, Sarvajanik University, Surat, India
e-mail: mayuri.mehta@scet.ac.in

M. Mehta et al. (eds.), *Explainable AI: Foundations, Methodologies and Applications*, Intelligent Systems Reference Library 232,
https://doi.org/10.1007/978-3-031-12807-3_3

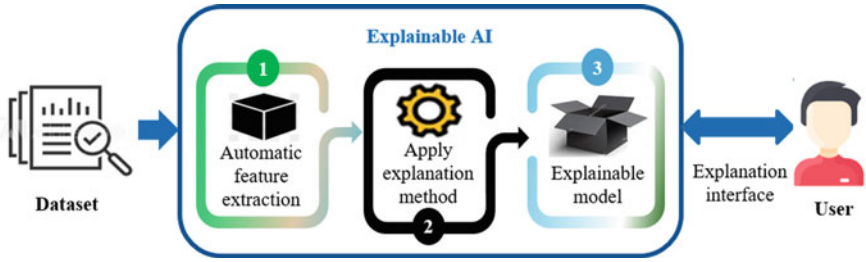


Fig. 3.1 Pipeline of an explainable AI method

black-box model made a particular decision. Moreover, it helps the researcher retrace the model functionality for deducing the inference mechanism. The methods that explain and interpret the internal mechanism and the decisions of machine learning and deep learning models are known as explanation methods. These methods explain the model decision and logic in different forms of explanation. Figure 3.1 illustrates a general pipeline of an explainable AI.

Interpretability is another term commonly used as a synonym for explainability. However, interpretability is the capability of a model to provide inference in an understandable form. An interpretable model enables end-user to prevent model bias, get the feature importance on model output, test reliability and ultimately assist in debugging the model. The complexity of a machine-learning model is directly related to its interpretability. Generally, complex models have higher accuracy, but they are more difficult to interpret and explain because they are developed with many hypermeters. Furthermore, the interpretability of machine learning and deep learning models often compromises the accuracy and overall performance of the models. Thus, the most straightforward way to get to an interpretable model is to design an intrinsically interpretable model.

The model explanation is critical when an AI-based system generates an unexpected result. The explanation method enables control over a model, which helps identify and rectify flaws and unknown biases. Moreover, the explanation method justifies the decision made by a model to answer why a particular decision was made (Failed 2022; Adadi and Berrada 2018). When the user understands how a model produces the result, they can easily improve the model to enhance its capability. The user understanding of the model boosts users' trust in critical decision-making using an AI-based system. It also enhances trust and human acceptability toward AI-based solutions (Inam et al. 2021; Linardatos et al. 2021). Additionally, the model explainability enables an AI-based system to add new knowledge to its knowledge base and learn new rules or behaviour for smart decision making. As per literature, the explainability of AI systems is needed mainly for four reasons: (1) to justify decisions, (2) to control the inner working of black box models, (3) to improve the model to produce expected results, and (4) to build the trust of human on model results. Figure 3.2 illustrates the reasons and needs for explainable AI.

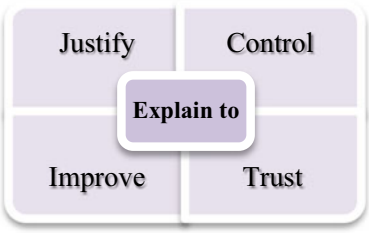


Fig. 3.2 Need for explanation in AI-based system

The research community introduced a variety of explanation methods and used them to produce understandable results from the black box model. This chapter aims to give a concise overview of different forms of explanation, existing XAI methods and frameworks that help researchers select the best possible method as per their application. The rest of the chapter is organized as follows: Sect. 3.2 presents the proposed classification of methods suitable for XAI and their brief description. Section 3.3 describes different forms of explanation useful for XAI. Section 3.4 presents a review of the six most popular XAI frameworks. Finally, the chapter is concluded in Sect. 3.5. In addition, our observations and future directions are discussed in this section.

3.2 XAI Methods and Their Classifications

This section discusses techniques and approaches used to explain black box models. It's important to incorporate explainability for better trust and understanding in black box models (Inam et al. 2021). Numerous XAI methods have been developed to explain the inner working of black-box models and their predictions (Linardatos et al. 2021). As shown in Fig. 3.3, we classify these methods in four different ways by considering different aspects of each method: (1) based on the scope of explainability, (2) based on implementation, (3) based on applicability, and (4) based on explanation level. This classification helps select suitable methods for explaining and interpreting black box models.

3.2.1 Based on the Scope of Explainability

The scope of explanation is one of the ways to classify explainable AI methods. It characterizes the extent of an explanation generated by an XAI method. An explanation can either describe the entire model or partially describe the model based on individual input instances (Mohseni et al. 2021). According to the scope of explainability, the explanation can be global or local (Ras et al. 2018; Doran et al. 2018; Adadi and Berrada 2018; Doshi-Velez and Kim 2017).

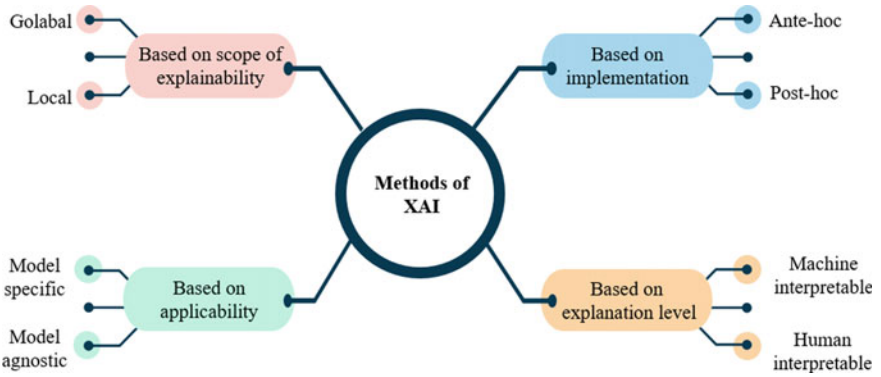


Fig. 3.3 Classification of explanation methods

The global methods are used for the explainability of the whole model. They provide a global explanation for the model's inner workings and decision-making (Ibrahim et al. 2019). A global method approximates the overall behaviour of a model to produce general representations of the model's relationship with its input instances. Symbolic representation is a commonly used technique to generate an interpretable representation of all predictions made by the model (Confalonieri et al. 2021). Bayesian rules and generalized additive models are a few examples of the global explanation method (Alicioglu and Sun 2022).

Unlike the global method, local methods are used to explain and understand individual predictions of a model. Explanations are built by generating local surrogate models considering individual prediction and input instances. The surrogate models are intrinsically interpretable models used to explain a complex model. They are trained on the predictions of the black box model to generate an explainable model. Local explanation methods often use saliency methods (Adadi and Berrada 2018; Linardatos et al. 2021; Angelov et al. 2021) to explain the relationship between specific input–output pairs (Mohseni et al. 2021). However, the local methods' explanations vary greatly depending on the instance considered (Confalonieri et al. 2021). Local methods like Local Interpretable Model-Agnostic Explanations (LIME) (Palatnik de Sousa et al. 2019), Shapley Additive Explanations (SHAP) (Failed 2019) and Deep Learning Important FeaTures (DeepLIFT) provide local explanations for an instance of a model. Generating model explanations using the local method is easier than explainability using the global method.

3.2.2 Based on Implementation

The explanation of the model can be generated during the training of a model, or it can be generated after model creation. Based on the way the explainability of the model

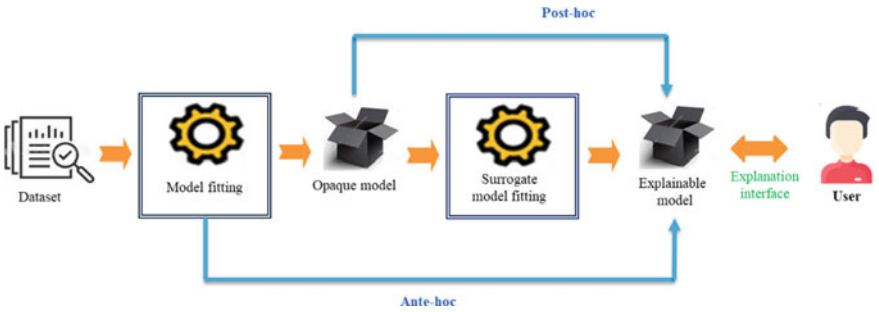


Fig. 3.4 Working of Ante-hoc and post-hoc methods

is implemented, XAI methods are categorized into two categories: ante-hoc and post-hoc methods. Figure 3.4 illustrates the working of ante-hoc methods and post-hoc methods. Ante-hoc methods are used to generate explanations from the beginning of the model training. They incorporate explanation directly into the structure of a model during design (Holzinger et al. 2017). Ante-hoc method, also known as the intrinsic method, often uses glass-box approaches for intrinsic explanations of models (Holzinger et al. 2019). Ante-hoc methods use decision trees or rules to explain how a prediction has been made through the model parameters. The models generated using ante-hoc methods are transparent and self-explanatory models such as bayesian rules and tree-based models (Islam et al. 2022).

Post-hoc explanation methods are used for models that are not readily explainable by their design (Holzinger et al. 2017; Barredo Arrieta et al. 2020a). They explain the inner working and inference mechanism of an already developed model or a newly created model after completing its training process. Generally, post-hoc methods mimic the behaviour of an already developed model to an external explainable model (Islam et al. 2022). Grad-CAM, Layer-wise Relevance Propagation (LRP), LIME, and Saliency Maps (Adadi and Berrada 2018; Linardatos et al. 2021; Angelov et al. 2021) are the most common examples of post-hoc methods.

3.2.3 Based on Applicability

Based on the application of explanation on different models, explanation methods are further categorized into two categories: model specific and model agnostic. The model agnostic methods are applied to any model, whereas the model-specific methods are restricted to particular models (Islam et al. 2022). Model-specific explanations are intrinsic methods where explanations are limited to a specific class of model. These methods aim to bring transparency to a particular type of model.

Model agnostic methods are used to interpret already developed models (Failed 2019). The advantage of these methods is that they do not impact the model's performance because they are independent of the inner working of the model (Linardatos

et al. 2021; Dosilovic et al. 2018). These methods have been often used due to their flexibility in the architecture of a model. Model agnostic methods also provide post-hoc explanations. LIME, LRP, and SHAP are popular examples of model-agnostic post-hoc explainers (Alicioglu and Sun 2022; Zhang et al. 2022). Following are some post-hoc model agnostic methods to achieve an understanding of a model.

(i) Attribution method: The attribution method analyzes the sensitivity of how the output is influenced by its input and/or weight perturbations. In other words, it measures the importance of each attribute or feature to the prediction. Sensitivity analysis, LRP, and feature importance are examples of the attribution method (Adadi and Berrada 2018; Chakraborty et al. 2018).
(ii) Visualization method: Visualization method explores the pattern hidden inside a learned model by visualizing its representations (Ras et al. 2018; Kim et al. 2019; Rajaraman et al. 2018). These techniques are essentially used for supervised learning models. Surrogate models, partial dependence plots and individual conditional expectations are examples of Visualization methods. Partial dependence plots are used to identify relationships between a set of features with the model outcome.
(iii) Knowledge extraction: The knowledge extraction method provides a comprehensible description of the knowledge by approximating the decision-making process using the input and output of the given model. Either rule extraction or model distillation can be used to gain insight into the model.

3.2.4 Based on Explanation Level

The explanations of XAI models are either analyzed by a machine (i.e. bot) or presented to a human. In machine-interpretable methods, users can mathematically analyze algorithmic mechanisms of predictions made by the system. This type of explanation is known as machine-to-machine explainability. Machine reasoning explainability uses techniques like compositionality, computational argumentation, and iterative contrastive explanations (Inam et al. 2021).

Despite the advantages of machine-interpretable methods, their explanations fail to generate human-understandable models (Confalonieri et al. 2021; Angelov et al. 2021). Human interpretable methods use linguistic variables and symbols to provide user-centric explanations of how a decision has been made (Holzinger et al. 2017; Chakraborty et al. 2018; Abdul et al. 2018). Rationalization is a form of explanation that attempts to explain a model prediction based on how a human would explain it. AI rationalization closely resembles explanations that are most likely given by a human (Ehsan et al. 2018).

3.3 Forms of Explanation

This section presents the proposed classification of different forms of explanation suitable for XAI. Many types of explanations are generated and used to explain the predictions made by machine learning and deep learning models. A broader range of end-users, such as naive users, data scientists and domain experts with different perspectives, demand the explanation in their understandable terms. Hence, selecting the best form of explanation suitable for a given model is a prime concern. Our literature study identified four different forms of explanations commonly used for deducing a decision (Ibrahim et al. 2019; Islam et al. 2022). Figure 3.5 illustrates these four forms of explanations: (1) analytical explanation, (2) visual explanation, (3) rule-based explanation, and (4) textual explanation. The analytical explanations are appropriate for domain experts and data scientists. Visual and rule-based explanations are found to be suitable for naive users. The textual explanations are presented in natural language for the general users.

3.3.1 Analytical Explanation

Analytical explanations are generated by measuring the contribution of the input features to the model's outcome. They are represented by various numeric metrics such as saliency, causal importance, feature importance, features confidence score, and mutual importance (Islam et al. 2022). Domain experts mostly utilize them to view and explore the data concerning their feature importance. They are suitable for post-hoc methods adopted for already developed models. Confident itemsets explanations presented in Moradi and Samwald (2021) are among the best examples of analytical explanation. The confident itemsets explanations utilize confidence score to find confident itemsets by considering each word as an item in the text record. Figure 3.6 illustrates an example of confident itemsets explanation for a text record "Where is mile high stadium?" from the TREC question classification dataset (Moradi and Samwald 2021).

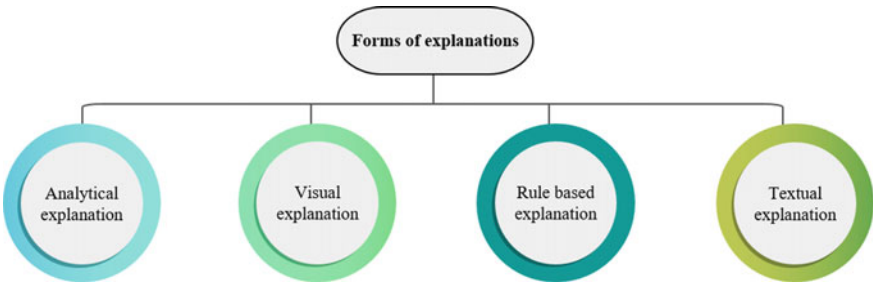


Fig. 3.5 Different forms of explanation useful for XAI

Fig. 3.6 Example of analytical explanation

Text record: Where is mile high stadium? **Prediction**: *LOC: other*			
Explanation using confident itemsets explanations: Minimum confidence threshold: 0.6			
Class: LOC: other **Score:** 2.554		**Class:** NUM: count **Score:** 0.666	
Itemset	**Confidence**	**Itemset**	**Confidence**
<where>	0.888	<mile>	0.666
<stadium>	0.666		
<where>, <stadium>	1.0		

3.3.2 Visual Explanation

Visual explanations are one of the most commonly used explanation forms for computer vision tasks. Visual explanations use visualization techniques such as class activation maps, gradient-based class activation maps (Abdul et al. 2018; Guidotti et al. 2018; Kim et al. 2018; Yasaka and Abe 2018) and attention maps to explain the model's prediction (Alicioglu and Sun 2022; Angelov et al. 2021). In these methods, the saliency heatmaps (Mohseni et al. 2021; Alicioglu and Sun 2022; Islam et al. 2022) using visual elements are generated to specify important regions of the input image. Visual explanations are post-hoc methods used for both the local and global explanations. Naive users can easily interpret visual explanations, which contain charts, trend lines, etc. An example of a visual explanation proposed by the authors of Sun et al. (2020) for fault diagnostics in an industrial machine is illustrated in Fig. 3.7. The bottom side nut of the machine is the most vibrating region because it is the most significant part of the fault diagnosis of the water pump.

Class Activation Map (CAM) is an explanation method suitable for convolutional neural networks with a global average pooling (Zhou et al. 2016). It highlights the discriminative region of the input image to identify the class of the image. It uses activation maps of the last convolutional layer to train linear classifiers for each class for final class estimation. The image regions that play an important role in prediction are identified by projecting the weights of the output layer onto the convolutional feature map. CAM requires retraining linear classifiers, one for each class, making this method time-consuming.

The authors of Selvaraju et al. (2016) have proposed Gradient-weighted CAM (Grad-CAM) to reduce the time complexity of CAM. Grad-CAM is the general form of CAM suitable for any convolutional neural network architecture. It uses the gradients of the target feature, flowing into the final convolutional layer to visualize the class activation maps. Moreover, it highlights the important regions in the input

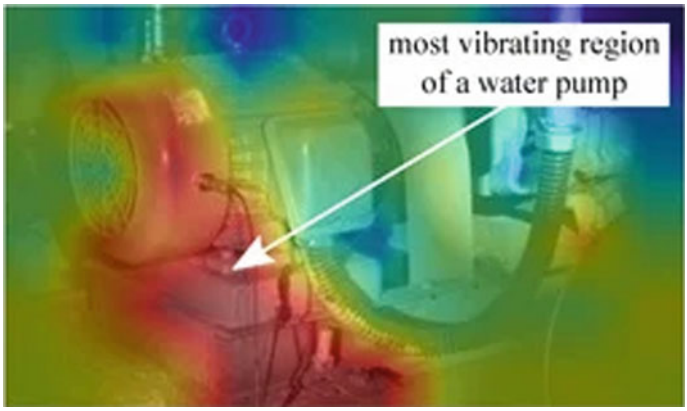


Fig. 3.7 Example of visual explanation (Sun et al. 2020)

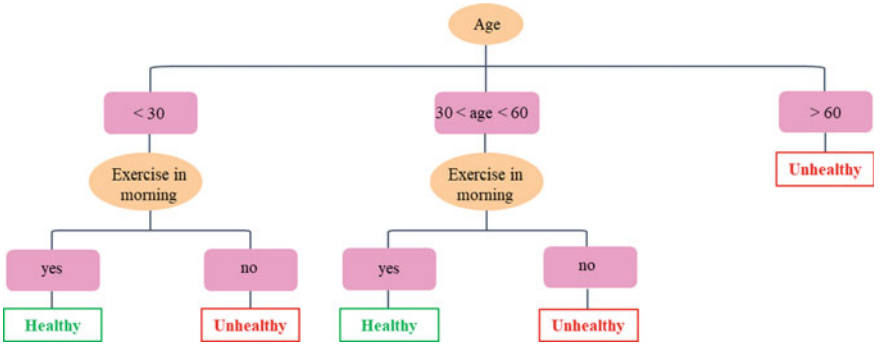
image for the selected feature. However, Grad-CAM fails in the localization of objects with multiple occurrences of the same class. The author of Selvaraju et al. (2016) also proposed a variation of Grad-CAM known as Guided Grad-CAM to obtain fine-grained pixel-scale representation. Guided Grad-CAM upsamples and fuses class-specific saliency map with the visualizations generated by guided backpropagation.

Grad-CAM++ (Chattopadhay et al. 2018) is an extension of the Grad-CAM that provides better visual explanations for the convolutional neural network. It enhances the object localization capability by extending multiple object instances in a single image. It utilizes second-order gradients of feature maps from the last convolutional layer. Additionally, it uses a specific class score as a weight to generate a visual explanation for the corresponding class label. The importance of each pixel is captured separately in the gradient feature map by assigning different weights to each pixel. Grad-CAM++ is more suitable for multi-label classification problems.

3.3.3 Rule-Based Explanation

The rule-based explanations are the simplest form of explanation. They describe the model inference mechanism using a set of IF–THEN rules or a tree (Islam et al. 2022). An example of a rule-based explanation along with a decision tree is illustrated in Fig. 3.8. In this example, two significant attributes, *age* and *exercise in the morning,* are considered to predict whether a person is healthy or unhealthy. Most rule-based explanations are ante-hoc methods that interpret models with a global scope. Ensemble learning and decision tree are popular examples of rule-based explanations. This type of explanation is commonly used to develop recommendation systems for naive users.

Adaptive Neuro Fuzzy Inference System (ANFIS) (Sagir and Sathasivam 2017; Keneni et al. 2019; Mehrdad Aghamohammadi and Madan 2019) is another popular



IF-THEN rules

Is a person healthy?

IF age < 30 AND Exercise in morning = yes, THEN Healthy.

IF age < 30 AND Exercise in morning = no, THEN Unhealthy.

IF 30 < age < 60 AND Exercise in morning = yes, THEN Healthy.

IF 30 < age < 60 AND Exercise in morning = no, THEN Unhealthy.

IF age > 60, THEN Unhealthy.

Fig. 3.8 Example of rule-based explanation

rule-based inference system that combines fuzzy inference rules with a neural network. Fuzzy Inference Systems (FIS) is a mapping method of a given input to output using the theory of fuzzy sets. There are two broad categories of FIS, namely the Mamdani fuzzy system and the Sugeno fuzzy system (Keneni et al. 2019). The general structure of the Mamdani fuzzy system is as follows:

IF p is A and q is B, THEN z is C;

where p and q are linguistic input variables and Z is the linguistic output.

Similarly, in the case of the Sugeno fuzzy system, the rules format is as follows:

IF p is A and q is B, THEN z is f(p,q);

where p and q are fuzzy sets on a universe of discourse and z is an output in the form of mathematical function f(p,q).

3.3.4 Textual Explanation

The textual explanations present the model prediction by learning text explanations as sets of words denoting the features that influence the model prediction (Bennetot

et al. 2019). They use Natural Language Processing (NLP) techniques to describe the model prediction in natural language. Moreover, they utilize numerous methods for generating symbols that represent the inner working of a model (Barredo Arrieta et al. 2020b). However, textual explanations are the least common among all forms of explanations due to their high computational requirement for NLP tasks. Generally, textual explanations are generated for individual prediction for precise and specific results. They are suitable for the general user. They are popular in applications like interactive question-answering systems (Mohseni et al. 2021). An example of textual explanation is given below based on Fig. 3.8.

Example of textual explanation: "*The person is classified as 'unhealthy' RATHER THAN 'healthy' because person age is more than 30 and no exercise in morning*".

The textual explanations are based on either factors or features that influence the model prediction or representative examples that support the prediction. If explanations are constructed for humans, they should be contrastive or counterfactual (Stepin et al. 2021; Zucco et al. 2018). Several researchers emphasize that good explanations are contrastive that explain the "Why", "Why not", and "What-if" of an AI-based system. The contrastive explanation is an effective method for mental model formation. Moreover, contrastive explanations improve the understanding without providing a full causal analysis (Kim et al. 2016). The counterfactual explanations are used to explain predictions of individual instances (Myers et al. 2020). A counterfactual explanation is a human-friendly explanation that describes a causal situation in the form: If an event "P" had not occurred, "Q" would not have occurred. Additionally, the counterfactual explanation identifies external factors that affect the model output.

3.4 Frameworks for Model Interpretability and Explanation

Researchers have designed several state-of-the-art frameworks to develop interpretable machine learning models. This section discusses the six promising XAI frameworks developed in python. These six frameworks are selected based on their explanation generation capability, application, and success on standard AI systems.

3.4.1 Explain like I'm 5

Explain Like I'm 5 (ELI5) is a python framework that helps to debug machine learning and deep learning models. ELI5 is one of the simple frameworks that finds the importance of each feature to the output for understanding the inner working of a model. However, its explanation is limited to parametric linear models and decision tree-based models. ELI5 provides two major functions: eli5.show_weights() function to inspect model parameters and eli5.show_prediction() function to inspect an individual prediction and determine why the model predicts this.

3.4.2 Skater

Skater is another popular open-source python framework designed to understand the inner workings of the black box model. The Skater framework evaluates and explains predictive models based on independent (input) and dependent (target) variables in a post-hoc manner (Linardatos et al. 2021). Moreover, it enables better model insight and debug options by keeping humans in the loop. The Skater framework supports a variety of models which can be explained either at the local or global level. In addition, it also supports object-oriented and functional programming paradigms to provide better scalability and parallelism.

3.4.3 Local Interpretable Model-Agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) (Palatnik de Sousa et al. 2969) is a surrogate-based explanation method that explains a model's prediction by fitting a local surrogate model whose predictions are easy to explain. LIME explains each prediction to understand how the black box model works in that local fidelity. It observes the effects of individual predictions by perturbing the original data. Although LIME is popular and simple, random perturbation of LIME results in unstable interpretation results. The authors of Zafar and Khan (2021) have proposed a deterministic version of LIME known as DLIME to deal with this limitation. Unlike LIME, DLIME uses hierarchical clustering to group the data and k-nearest neighbours to find the cluster where the given instance belongs.

3.4.4 Shapley Additive Explanations

Shapley Additive Explanations (SHAP) is used to explain an instance's prediction by computing each feature's contribution to the prediction. Shapely values are used to

identify the effect of individual features on the model outcome (Failed 2019). Shapley values provide explanations by assigning a value called weight to each feature for a particular prediction. SHAP can guarantee consistency and local accuracy because of its thorough approach to considering all possible predictions, such as using all possible combinations of inputs. SHAP is available in two variants (i) KernelSHAP (Lundberg and Lee 2017) and (ii) TreeSHAP (Linardatos et al. 2021). Kernel SHAP is a model agnostic method based on LIME concepts and Shapley values. The major drawback of Shapley values is their computational complexity. Tree SHAP computes exact SHAP values for decision trees based models. Asymmetric Shapley Values (ASV) is a SHAP variation that incorporates a causal graph of the cause-effect relationship between variables in the model explanation process. Unlike SHAP, where shapely values are symmetrical, ASV uses asymmetric shapely values. The model fairness analysis is a major application of ASV values because it can capture the indirect effects of the variable on a model.

3.4.5 Anchors

The anchors method explains each model prediction by finding IF–THEN rules called anchors (Ribeiro et al. 2018). An anchor explanation is a rule framed using input and the model prediction at a local level. Anchors are high precision and model-agnostic explanation methods that use reinforcement learning to construct rules without knowledge about the model. Moreover, they can explain nonlinear models because they work on feature predicates. The key limitation of Anchors is that they only support textual and tabular data. They can produce explanations in the form of tabular data and text depending on the application domain.

3.4.6 Deep Learning Important Features

Deep Learning Important FeaTures (DeepLIFT) is another popular explanation framework for the deep neural network. It calculates the importance score of each feature. It explains the model by computing the difference in model output from some reference output based on the difference of the input from some reference input. The reference input represents some default input (Shrikumar et al. 2017). Moreover, DeepLIFT provides different considerations for positive and negative contributions.

The aforementioned XAI frameworks are critically examined and their comparative analysis is presented in Table 3.1. All six frameworks support model agnostic post-hoc explanation at the local level. In addition to local explanation, SKATER and SHAP support the global explanation.

Table 3.1 Comparison of explainable framework

Explainable framework	Method			Form of explanation supported
	Local/global	Post-hoc/atte-hoc	Model agnostic/model specific	
ELI5	Local	Post-hoc	Model agnostic	Textual
SKATER	Local and global	Post-hoc	Model agnostic	Textual
LIME	Local	Post-hoc	Model agnostic	Textual and visual
SHAP	Local and global	Post-hoc	Model agnostic	Textual and visual
ANCHORS	Local	Post-hoc	Model agnostic	Textual and tabular
DeepLIFT	Local	Post-hoc	Model agnostic	Textual

3.5 Conclusion and Future Directions

While Artificial Intelligence has a long history, Explainable AI, also known as XAI, is a relatively new interdisciplinary research field. XAI research has been growing rapidly due to the increasing demand for diverse frameworks and methods to produce interpretable and understandable results from black box models. Many methods of Explainable AI have been proposed in the literature to understand the inner working of black box models and their predictions. This chapter provides a selective and summarized overview of various methods and forms of explanation for XAI. A taxonomy of methods suitable for XAI followed by a classification of explanation forms has also been proposed. Moreover, a comparative analysis of six popular XAI frameworks has been presented to help the research community select a suitable explainable framework.

XAI brings significant benefits to many application domains relying on AI-based systems. The potential application domains such as healthcare, finance, military, criminal justice and transportation require more attention to the human's role in existing explainability methods because the consequence of decisions can be dangerous. It has been observed that little attention has been given to combining different interpretability methods to achieve easy to understand and human-centric explanations. Hence, developing general and interactive explanations methods with emerging NLP techniques will be a new research direction in XAI.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of Conference on Human Factors in Computing Systems, pp. 1–18 (2018). https://doi.org/10.1145/3173574.3174156

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

Aghamohammadi, M., Madan, M., Hong, J.K., Watson, I.: Predicting heart attack through explainable artificial intelligence. In: International Conference on Computational Science—ICCS 2019, vol. 1, pp. 633–645 (2019). https://doi.org/10.1007/978-3-030-22741-8

Alicioglu, G., Sun, B.: A survey of visual analytics for Explainable Artificial Intelligence methods. Comput. Graph. **102**, 502–520 (2022). https://doi.org/10.1016/j.cag.2021.09.002

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **11**(5), 1–13 (2021). https://doi.org/10.1002/widm.1424

Barredo Arrieta, A. et al.: Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion (2020a). https://doi.org/10.1016/j.inffus.2019.12.012

Barredo Arrieta, A. et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020b). https://doi.org/10.1016/j.inffus.2019.12.012

Bennetot, A., Laurent, J.L., Chatila, R., Díaz-Rodríguez, N.: Towards explainable neural-symbolic visual reasoning, *arXiv Learn.* (2019)

Chakraborty S. et al.: Interpretability of deep learning models: a survey of results (2018). https://doi.org/10.1109/UIC-ATC.2017.8397411

Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097

Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable Artificial Intelligence, *Wiley Interdiscip.* Rev. Data Min. Knowl. Discov. **11**(1), 1–21 (2021). https://doi.org/10.1002/widm.1391

Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. In: CEUR Workshop Proceedings, vol. 2071 (2018)

Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning, *arXiv Prepr. arXiv1702.08608*, no. Ml, pp. 1–13 (2017). http://arxiv.org/abs/1702.08608

Dosilovic, F.K., Brcic, M., Hlupic, N.: Explainable artificial intelligence: a survey. In: Proceedings of 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, pp. 210–215 (2018). https://doi.org/10.23919/MIPRO.2018.8400040

Ehsan, U., Harrison, B., Chan, L., Riedl, M.O.: Rationalization: a neural machine translation approach to generating natural language explanations. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 81–87 (2018). https://doi.org/10.1145/3278721.3278736

Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of 2018 IEEE 5th International Conference on Data Science Advanced Analytics DSAA 2018, pp. 80–89, (2019). https://doi.org/10.1109/DSAA.2018.00018.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–45 (2018). https://doi.org/10.1145/3236009

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **9**(4), 1–13 (2019). https://doi.org/10.1002/widm.1312

Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? No. Ml, pp. 1–28 (2017). http://arxiv.org/abs/1712.09923

Ibrahim, M., Louie, M., Modarres, C., Paisley, J.: Global explanations of neural networks: mapping the landscape of predictions. *CoRR arXiv1902.02384*, pp. 1–10 (2019). http://arxiv.org/abs/1902.02384

Inam, R., Terra, A., Mujumdar, A., Fersman, E., Feljan, A.V.: Explainable AI—how humans can trust AI. Ericsson, no. April, pp. 1–22, 2021. https://www.ericsson.com/En/Reports-and-Papers/White-Papers/Explainable-Ai--How-Humans-Can-Trust-Ai

Islam, S.R., Eberle, W., Ghafoor, S.K., Ahmed, M.: Explainable artificial intelligence approaches: a survey. CoRR, pp. 1–14 (2021). http://arxiv.org/abs/2101.09429

Islam, M.R., Ahmed, M.U., Barua, S., Begum, S.: A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. **12**(3) (2022). https://doi.org/10.3390/app12031353

Keneni, B.M. et al.: Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. IEEE Access, vol. 7, no. c, pp. 17001–17016 (2019). https://doi.org/10.1109/ACCESS.2019.2893141

Kim, B., Khanna, R., Koyejo, O.: Examples are not enough, learn to criticize! Criticism for interpretability. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 2288–2296 (2016)

Kim, I., Rajaraman, S., Antani, S.: Visual interpretation of convolutional neural network predictions in classifying medical image modalities. Diagnostics (2019). https://doi.org/10.3390/diagnostics9020038

Kim, B. et al.: Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: 35th International Conference on Machine Learning, ICML 2018, vol. 6, pp. 4186–4195 (2018)

Krajna, A., Brcic, M.: Explainable artificial intelligence : an updated perspective explainable artificial intelligence : an updated perspective. (2022)

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 1–45 (2021). https://doi.org/10.3390/e23010018

Lundberg, S., Lee, S.-I.: A unified approach to interpreting model predictions. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), May 2017, pp. 1–10. http://arxiv.org/abs/1705.07874. Accessed 30 Aug 2019

Messalas, A., Kanellopoulos, Y., Makris, C.: Model-agnostic interpretability with Shapley values. In: 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), July 2019, pp. 1–7. https://doi.org/10.1109/IISA.2019.8900669

Mohseni, S., Zarei, N., Ragan, E.D.: A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Trans. Interact. Intell. Syst. **11**(3–4), 1–45 (2021). https://doi.org/10.1145/3387166

Moradi, M., Samwald, M.: Post-hoc explanation of black-box classifiers using confident itemsets. Expert Syst. Appl. **165**, 113941 (2021). https://doi.org/10.1016/j.eswa.2020.113941

Myers, C.M., Freed, E., Pardo, L.F.L., Furqan, A., Risi, S., Zhu, J.: Revealing neural network bias to non-experts through interactive counterfactual examples (2020). http://arxiv.org/abs/2001.02271

Palatnik de Sousa, I., Maria Bernardes Rebuzzi Vellasco, M., Costa da Silva, E.: Local interpretable model-agnostic explanations for classification of lymph node metastases. Sensors **19**(2969), 1–18 (2019). https://doi.org/10.3390/s19132969

Rajaraman, S., Candemir, S., Kim, I., Thoma, G., Antani, S.: Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. Appl. Sci. (2018). https://doi.org/10.3390/app8101715

Ras, G., Van Gerven, M., Haselager, P.: Explanation methods in deep learning: users, values, concerns and challenges, pp. 19–36 (2018). https://doi.org/10.1007/978-3-319-98131-4_2

Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: high-precision model-agnostic explanations. In: 32nd Proceedings of the AAAI Conference on Artificial Intelligence, AAAI 2018, vol. 32, no. 1, pp. 1527–1535 (2018)

Sagir, A.M., Sathasivam, S.: A novel adaptive neuro fuzzy inference system based classification model for heart disease prediction. Pertanika J. Sci. Technol. **25**(1), 43–56 (2017)

Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization, *CoRR*, vol. abs/1610.0 (2016). http://arxiv.org/abs/1610.02391

Shi Zhang, Q., Chun Zhu, S.: Visual interpretability for deep learning: a survey. Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018). https://doi.org/10.1631/FITEE.1700808

Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences, *CoRR*, vol. abs/1704.0 (2017). http://arxiv.org/abs/1704.02685

Stepin, I., Alonso, J.M., Catala, A., Pereira-Farina, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021). https://doi.org/10.1109/ACCESS.2021.3051315

Sun, K.H., Huh, H., Tama, B.A., Lee, S.Y., Jung, J.H., Lee, S.: Vision-based fault diagnostics using explainable deep learning with class activation maps. IEEE Access **8**, 129169–129179 (2020). https://doi.org/10.1109/ACCESS.2020.3009852

Yasaka, K., Abe, O.: Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Med. **15**(11), 1–4 (2018). https://doi.org/10.1371/journal.pmed.1002707

Zafar, M.R., Khan, N.: Deterministic local interpretable model-agnostic explanations for stable explainability. Mach. Learn. Knowl. Extr. **3**(3), 525–541 (2021). https://doi.org/10.3390/make3030027

Zhang, Y., Weng, Y., Lund, J.: Applications of explainable Artificial Intelligence in diagnosis and surgery. Diagnostics **12**(2) (2022). https://doi.org/10.3390/diagnostics12020237

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

Zucco, C., Liang, H., Di Fatta, G., Cannataro, M.: Explainable sentiment analysis with applications in medicine. In: Proceedings of—2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018, pp. 1740–1747 (2019). https://doi.org/10.1109/BIBM.2018.8621359