Chapter 4
Methods and Metrics for Explaining Artificial Intelligence Models: A Review

Puja Banerjee and Rajesh P. Barnwal

Abstract Deep learning (DL) solutions have been facing the long-standing problem of making the Explainable Artificial Intelligence (XAI) an integral part of the machine learning pipeline. In recent times, multiple deep learning approaches have been established for solving the enhanced complications aroused due to high predictive capacity. Though DL models demonstrate exceptionally high accuracy but the same comes with computationally complex and difficult to interpret black-box architectures. Several efforts are being made to develop the methods for making such high-precision black-box models explainable so that the trustworthiness and reliability of such models can be established. The chapter provides an overview of XAI, different methods of XAI, and metrics associated with those methods. Further, the chapter also discusses the motivational factors behind XAI, its applications, and its taxonomy. For clarity on the XAI implementation stage, Pre-model, In-model, and Post-model explainability are elaborated along with the model-agnostic and model-specific techniques. The chapter concludes with a brief discussion on a simple use-case of implementing the XAI method in a real-life problem followed by enumerating possible future research directions.

4.1 Introduction

Most effective Artificial Intelligence (AI) techniques are still black-box. The users and AI professionals are unable to interpret and understand the reason behind the decisions of those techniques. The absence of such a transparent system can result

P. Banerjee
Academy of Scientific and Innovative Research, Ghaziabad, India
e-mail: puja.cmeri20a@acsir.res.in

R. P. Barnwal (✉)
AI & IoT Lab, IT Group, CSIR-Central Mechanical Engineering Research Institute, Durgapur, India
e-mail: r_barnwal@cmeri.res.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_4

in severe consequences in the areas of health diagnosis, finance management, military, self-driving vehicles, etc. Thus, methods for explaining AI decision-making processes have experienced a huge increase of attention from the research community to its application domain. The methods of deep learning have advanced the state-of-the-art of Artificial Intelligence to the next level. However, even with such extraordinary improvements, the absence of explainations in the deep learning predictions and the lack of control over the internal functioning of model development act as major drawbacks. Thus efforts are focused on making deep learning models manageable and interpretable to the end-users. The International market of Explainable AI is sub-divided on the grounds of contributing better services and solutions. Depending on its domain of applications, the market of XAI is classified as drug discovery, fraud detection, advertising, recommendation engines, computer vision (e.g., in case of classifying images, visual question answering, image captioning), security, natural language processing (classification of text, sentiment analysis) as well as supply chain management, etc. Moreover, based on user industry, the XAI international market is classified into telecommunication, medicine, healthcare, retail, public sectors, logistics, entertainment, military, defense, etc. Based on region, the market of XAI is divided into North America, Europe, and the Asia Pacific. In the report published by Next Move Strategy Consulting (Next Move Strategy Consulting (NMSC) 2022), the international market of XAI is expected to generate an amount of $21.78 billion (USD) by 2030.

Deep Learning is a common and most contemporary subset of the wide Machine Learning which field of study utilizes Artificial Neural Networks(ANN) (Goodfellow et al. 2016) for prediction and inference. Though the advancement of ANN started between the 1940s and 1960s, researchers did not find sufficient computational power and data for testing the limits of the developed methods. Deep Learning uses deeper networks (more number of model layers) for prediction. The deep learning structure enabled non-linear modeling of the data for prediction, as opposed to the traditional paradigms that follow linear modeling for prediction. The term “deep”, in these methods is derived from the idea of using multiple layers in learning networks. The layers comprise of input and an output layer and multiple hidden layers between these two layers. On the other hand, shallow networks are networks that have up to two hidden layers. Deep Learning can achieve higher accuracy because of the multiple hidden layers which can better model higher levels of abstractions in the data. These models can also better leverage the data in big data sets for prediction. Since these models can efficiently leverage the information in big data sets (often containing millions of data points), the chance of an incorrect prediction due to the unavailability of a data point critical to the prediction gets significantly reduced. This mitigates the risk of inaccurate predictions and therefore leads to an overall improvement in the prediction accuracy.

4.1.1 *Bringing Explainability to AI Decision—Need for Explainable AI*

XAI is a field of research that makes AI-based prediction systems transparent and understandable to end-users. This term was first coined in the year 2004 (Adadi and Berrada 2018) for describing the capability of a system for explaining the behavior of entities managed by the AI models in simulating applications in games (Van Lent et al. 2004). Although the concept is still comparatively new, the difficulty of explaining the model has survived since the middle of the 1970s when researchers started studying explainations for the expert AI systems. Moreover, the advancement toward solving such problems has decelerated as AI reached a point of inflection with the extraordinary progress in Machine Learning techniques. Since then the center of research in the Artificial Intelligence domain has shifted towards model development and algorithm implementation. AI and Machine learning generally demonstrates practical success in many different application domains. Some of the popular applications sectors include autonomous vehicles and drones, speech recognition systems, face recognition, navigation system, social networking, etc. But the central problem of such models is that these models are not transparent and thus black-box. This means that even if the fundamental principles of the mathematical are well understood, they lack explicit interpretive knowledge representation. This results in increasing ethical, legal, and privacy issues. This in turn results in applying black-box approaches in personal, defense, and business operations that could be more difficult and untrustworthy due to their adverse implications. Thus it necessitates the need for explainable AI systems that can introduce more transparency and clarity. Based on the existing literature, the need for explaining the AI systems may stem from the reasons that the user should get a trustworthy, dependable, compliant, effective, fair, and robust decision as shown in Fig. 4.1.

Various factors for which explainability is required in AI models are as follows:

- **Explain to Justify**: Explaining an AI decision furnishes the information required to be justified, generally when the decision is unexpectedly made. This makes sure that an auditable and provable way is there for defending the algorithmic decisions for being ethical and fair, which would result in introducing trust in the AI decisions.
- **Explain to Control**: XAI is important not because of justifying any AI decision but also helps in preventing incorrect AI decisions. Thus, a better understanding of the behavior of AI systems provides significant visibility of unpredictable flaws and vulnerabilities. This helps in the rapid identification and correction of errors in critical situations, which in turn enables enhanced control over the system.
- **Explain to Improve**: Another cause for developing XAI models is to make continuous improvements. An AI model which can be understood and explained is the one to be easily improved. This is where the end-users get to know the reason why an AI system produces such specific outputs. This shows the possibility that how XAI could act as the base for ongoing improvement of man and machine interaction.

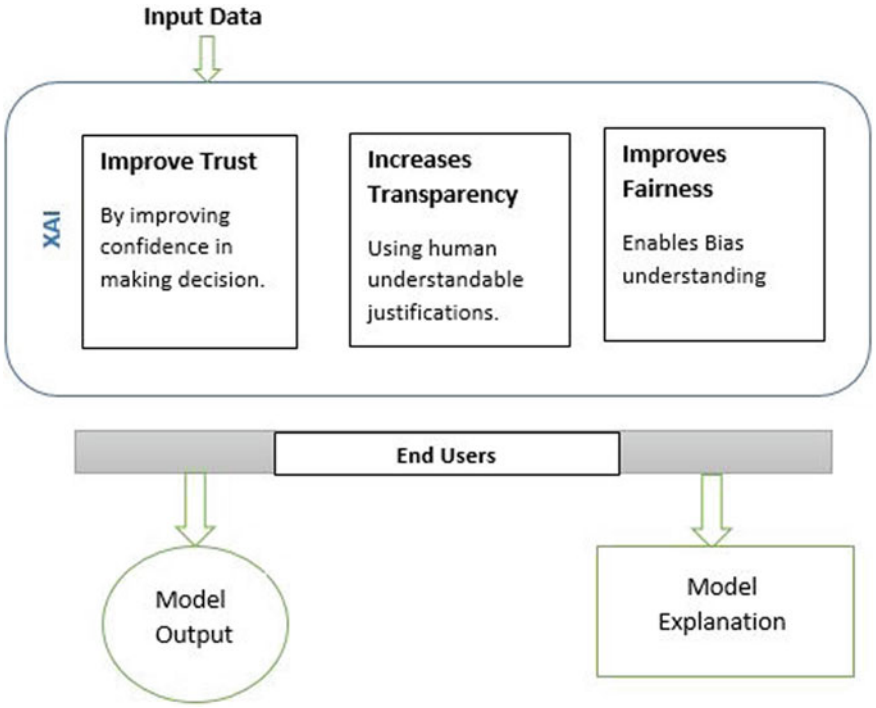


Fig. 4.1 Factors making XAI important

The key equivalent terminologies used in the literature for XAI and their interpretation are mentioned in Table 4.1.

Interpretability and explainability are approached via two categories, i.e, the transparency-based approach and the post-hoc-based approach. Transparency and estimated performance are contrary objectives and thus there should be a trade-off while developing an AI model. When a system is already robust and self-contained, introducing transparency is not necessary. But, when it is a part of another complicated system, then introducing transparency is good for debugging ability. Post-hoc interpretability takes out information from a trained model and does not specifically depend on the working of the model. An advantage of this type of approach is that it impacts the model's performance, which is treated as a black box.

4.2 Taxonomy of Explaining AI Decisions

Interests of researchers in the domain of Explainable AI are emerging. Previous works were concentrated on providing explainations of any decision undertaken based on knowledge-based expert AI systems. The main reason for these kinds of interest

Table 4.1 Equivalent key terminology used for XAI

Key terms	Description
Black-box AI	The AI model is a black box when it does not reveal anything about the internal design and structure (Suman et al. 2010) of the system. Because it is difficult for black box models to provide suitable explainations, the problem related to these systems is known as the *black-box problem*
Interpretable AI	Interpretable systems are those where the users can not only visualize the parameters necessary for any prediction but can also understand how the input variable is mathematically connected to the outputs. Researchers use the terms *interpretability* and *explainability* (Koh and Liang 2017) interchangeably. Others used terms such as *comprehensibility* or *understandability* (Bojarski et al. 2017) to refer to the same issue, whereas the term *interpretable AI* is more preferred in the industry
Responsible AI	This term of XAI takes societal, moral, and ethical values into consideration. The pillars of Responsible AI are Accountability, Transparency, and Responsibility (Dignum 2017)
Third-wave AI	Recently, the term third wave in AI has also surfaced, where the system constructs an explainatory model for classifying a real phenomenon and provides reason to their tasks and situations

in research related to XAI aroused because of current advancements in the area of Artificial Intelligence, it could be applied in a large range and into different domains, XAI is used in solving problems related to improper and unethical use, bias in any AI decisions, lack of transparency in the developed models. In addition to it, the latest laws imposed by the government illustrated that research on XAI should be done to a greater extent. XAI helps the AI systems in uncovering the hidden bias in any kind of decision caused due to the internal parameters of an opaque model.

A thorough review of different approaches was presented by Mueller et al. (2019) in the year 2019 which presented numerous types of explainations in AI systems and divided it into three generations i.e, the first generation -which includes efficient systems from the 70s. This system attempts in expressing the internal process of a system by putting knowledge directly from experts. The example includes changing the rules into natural expressions. The second-generation systems-these are intelligent systems, which are made to provide cognitive support to the human-machine

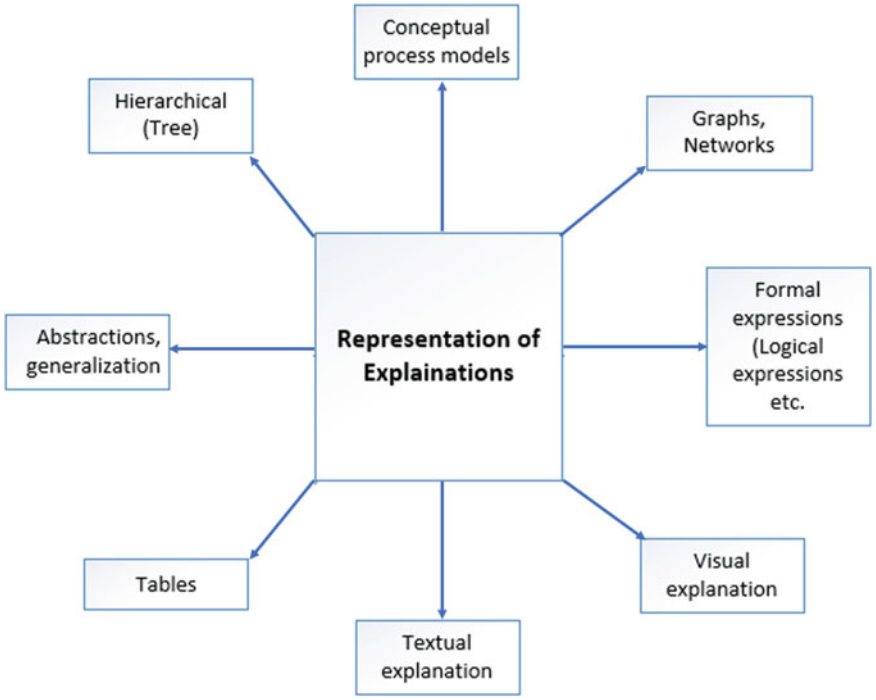


Fig. 4.2 Different representations for providing explainations for AI decisions

interaction and take the knowledge and reasoning capabilities of humans into consideration. For example, an interfacing system is being arranged in such a way that it complements the knowledge lacking by the users, and the third-generation systems, which utilizes techniques from the present era, started in the year 2015. Like first-gen systems, then disclose the internal working of the AI system. Although the systems are *black-box* in nature, it helps in clarifying their working. In addition to it, in recent days, researchers are using advanced computer technologies such as visualizing the data, and video animations, which have a huge potential in driving further research in XAI. Multiple ideas are being proposed for providing explainable decisions that aim in developing fair and trustable AI decisions.

Figure 4.2 illustrates different formats and representations of explaination. The concepts illustrated here show the *common ground* for research in the XAI domain. This is generated by properly reviewing and doing a thorough literature survey on AI.

For generating explaination, varieties of taxonomy are available, which include approaches that take notice of the type of explaination to be provided, a model which is needed to be explained, the scope for the explaination, or by combining all of these approaches. The taxonomy of Explainable Artificial Intelligence is shown in Fig. 4.3. Considering an example where XAI techniques have been applied the

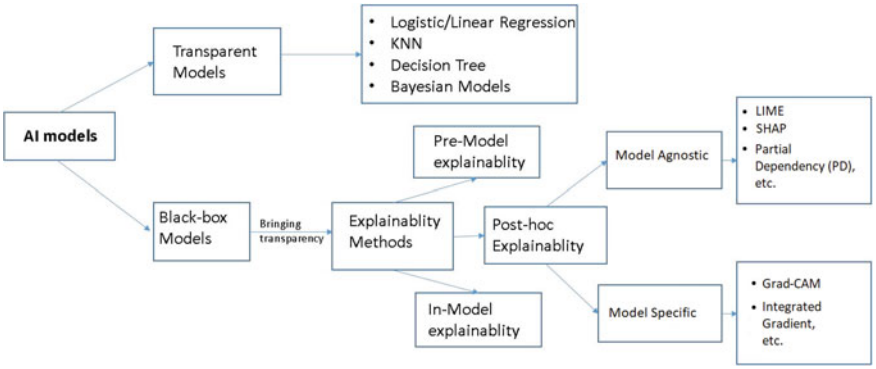


Fig. 4.3 Taxonomy of explainable artificial intelligence

methods for explaination are classified into an intrinsic and post-hoc method. In the intrinsic method, explainability is achieved by imposing constraints on the AI model. Post-hoc methods explain the model after its training. The post-hoc explainability method is further classified as Model-specific and Model-agnostic methods as well as local and global explainability methods. Moreover, Arya et al. (2019) showed a hierarchical relationship between different methods of explaination, the author also segregated explainations that are based on different used techniques and associated with different relations.

4.3 Methods of Explainable Artificial Intelligence

With improvements in hardware and a rise in data availability, the benefits of predictive performance increase using complex and opaque models. However, the primary issues that need to be addressed adequately are interpretability and explainability. With this kind of approach, we need to start our modeling with a trained black-box predictor and training data. While some methods deal with the issue of interpretability, others deal with explainability. The method is called model-agnostic when it operates on the inputs and outputs of the black-box model and is called model-specific when it uses the idiosyncrasies associated with some kind of representation.

These methods, in general, predict with an explaination, which is in the form of feature importance for a particular decision. Methods like layer-wise relevance propagation (Montavon et al. 2019) and sensitivity analysis (Christopher Frey and Patil 2002) have been presented to explain the predictions from DL models. Deep Taylor decomposition propagates explaination from Neural Network output to the input contribution. Model-agnostic methods are used to capture the degree of input influences on the system outputs. SHapley Additive exPlainations is a Local approximation method (Antwarg et al. 2019), which is used in explaining the prediction $f(y)$ for a single input y. SHAP is a framework that helps in estimating the number

of additive feature attributions which other multiple works follow in general. An explaination is generated by highlighting the relevant region of the input image.

The AI explaination is classified into three types:

- **Model-based explainations**: It represents those explainations that uses a model for explaining the original task (Krarup et al. 2019). In this type of explainations the task model explains itself or more interpretable is being generated to explain the task model.
- **Attribute-based explainations**: It generally measures the power of explaination of the input features and utilizes the ranks for explaining the task model (Jain et al. 2020). For instance, the feature importance explaination approach belongs to this class of explainations.
- **Example-based explainations**: This type of method selects instances from the testing and training data sets and can create new instances (van der Waa et al. 2021) for explaining the task model. For example, an instance is being selected which can be predicted well by the model as explainations, or by producing counterfactual examples for explainations.

Numerous explaination techniques have been described and evaluated which fall under following six broad approaches.

- **Feature relevance**: These approaches for explaining AI decisions focus on the inner functioning of the model and highlight features that best explain the output of the model.
- **Local explainations**: These approaches segment the solutions and provide explainations for smaller segments that are less complex. This tackles explainability by first dividing the solution and then generating explainations for less complex solution sub-spaces, which are relevant to the entire model. These types of explainations are given by those techniques which can only explain parts of the functionalities of the entire system, but not the complete whole.
- **Visualization**: These types of approaches allow end-users to visualize the behavior of the model, often by minimizing the complexity of the problems. These techniques of explainability aim to visualize the behavior of a model. Existing multiple visualization methods come with techniques that allow a simple human interpretable visualization approach. Visualizations can also be used with other kinds of methods for improving the understanding, and are considered the best way for explaining the complex interaction between the variables involved in the model to those users unfamiliar with AI modeling.
- **Explaination by example**: These approaches bring out specific representative data for explaining the overall behavior of the model.
- **Text explainations**: These types of approaches convert the explainations into natural language text. Text explaination generates symbols that represent the working of an AI model. These symbols can depict the algorithm rationale by using a semantic mapping technique that maps the model to the symbols.
- **Model simplification**: These approaches focus on building a new model that is less complex than a model that is to be explained.

4.3.1 Techniques of Explainable AI

- **Explainations by simplification**: These are the types of techniques that are used to explain an AI model. Because basic models are sometimes only representatives of particular parts in a model, this category includes local explainations of AI decisions. Nearly all the techniques, which are following this kind of path, working in simplifying the model, are based on some rule extraction methods. One of the familiar contributions to this kind of approach is the Local Interpretable Model Agnostic Explainations (LIME) (Ribeiro et al. 2016) method. This type of approach categorizes the explaination by simplifying it and provides local explainations. Along with LIME, there is another approach for the extraction of rules known as Genetic Rule Extraction. Even though it was not deliberated for extracting the rules from non-transparent models, the generic proposition of Genetic Rule Extraction is extended in accounting for explainability purposes. Explaining by Simplification is addressed from a separate perspective, where a technique for distilling and inspecting a black-box model is being presented. Within it, two major ideas are being illustrated: a technique for model distillation and comparison for auditing the risk of black-box scoring models; and a statistical test is being conducted for checking whether the auditing data is lacking the important features with which it was trained.
- **Feature relevance explainations**: These types of explaination methods are used for describing the working of an opaque model by ranking the feature and by calculating the influence, relevancy, and importance of each feature which results in predicting the output. A combination of propositions is being established inside this category, and they utilize different types of algorithmic approaches with a similar targeted goal. One of the important contributions to this is SHapley Additive exPlainations(SHAP) (Fidel et al. 2020). Researchers have introduced a method for calculating an importance score of the feature for every prediction with a set of necessary properties like consistency and local accuracy etc which is lacking by the antecedent. Other techniques for tackling the contribution of each feature for the prediction have an alliance with the game theory and local gradients.
- **Explaination by visualization**: This technique is used for achieving model agnostic explainations. Some of the works presented by Cortez et al. (2011) presented a detailed portfolio of the visualization method which helps in explaining a black box model made on some set of expanded techniques. Another set of visualization techniques presented by the same author (Cortez and Embrechts 2013) where three novel Stochastic-algebraic methods are Data-based Stochastic algebraic method, Monte-Carlo Stochastic algebraic method, and Cluster-based Stochastic algebraic method is being shown as one novel input measures. And finally, Goldstein et al. (2015) presented Individual Conditional Expectation plots as a mechanism for visualizing the estimating the model with the help of a supervised learning algorithm. Explaining the results visually is rare in the field of model-agnostic techniques. The design of these kinds of methods ensures that they can be easily applied to any kind of model without considering its inner structure, it creates

visualizations from the inputs and outputs of a non-transparent model, which is a difficult task. This is the reason nearly all the methods used for visualization fall into this type of categorical work together with the feature relevance technique, which provided the details that are finally presented to the end-user. Some of the important Explainability methods based on visualization techniques are as follows:

4.3.2 Stages of AI Explainability

The method of explainability could be classified in terms of the stages when this kind of method could be applied: before i.e., (pre-model), during the explaination i.e, (in-model), or after i.e., (post-model) when the AI model is built (Carvalho et al. 2019). Pre-modeling explainability methods are not dependent on the model, as because they are applied to the data itself. Different stages of Explainable AI are shown in Fig. 4.4. Pre-model explainability usually occurs before the model is selected, as because it is also important for exploring and having good knowledge about the data before the model is created. Pre-model explainability is, nearly related to data interpretability, which consists of explaining data analysis (Tukey et al. 1977) methods. Visualizing data is very critical in the case of pre-model explainability, which consists of a graphical representation of the data aiming to provide better knowledge . Model explainability deals with AI models which are inherently interpreting. Post-modeling explainability helps in improving interpretability after the model is built, thus it is referred to as post-hoc.

4.3.2.1 Pre-model Explainability Methods

Pre-model explainability is a collection of several strategies aimed at better comprehending the dataset used in model construction. Exploratory data analysis, dataset description standardization, explainable feature engineering, and dataset summarising methodologies are the four primary categories of pre-model explainability. These pre-model explainability methodologies are being discussed as follows:

- **Exploratory data analysis**: Exploratory data analysis aims to extract a summary of a dataset's primary properties. The mean, standard deviation, range, missing samples, and other statistical features of the dataset are frequently included in this summary. An exploratory data analysis task looks at the relative frequency of faulty and non-defective photographs which can show a problem with class imbalance, with considerably fewer defective photographs than non-defective ones. After identifying the problem in the training dataset, a variety of methods can be used to alleviate the problem and improve the classifier's performance. However, relying solely on statistical features is insufficient while analyzing data. Real-world

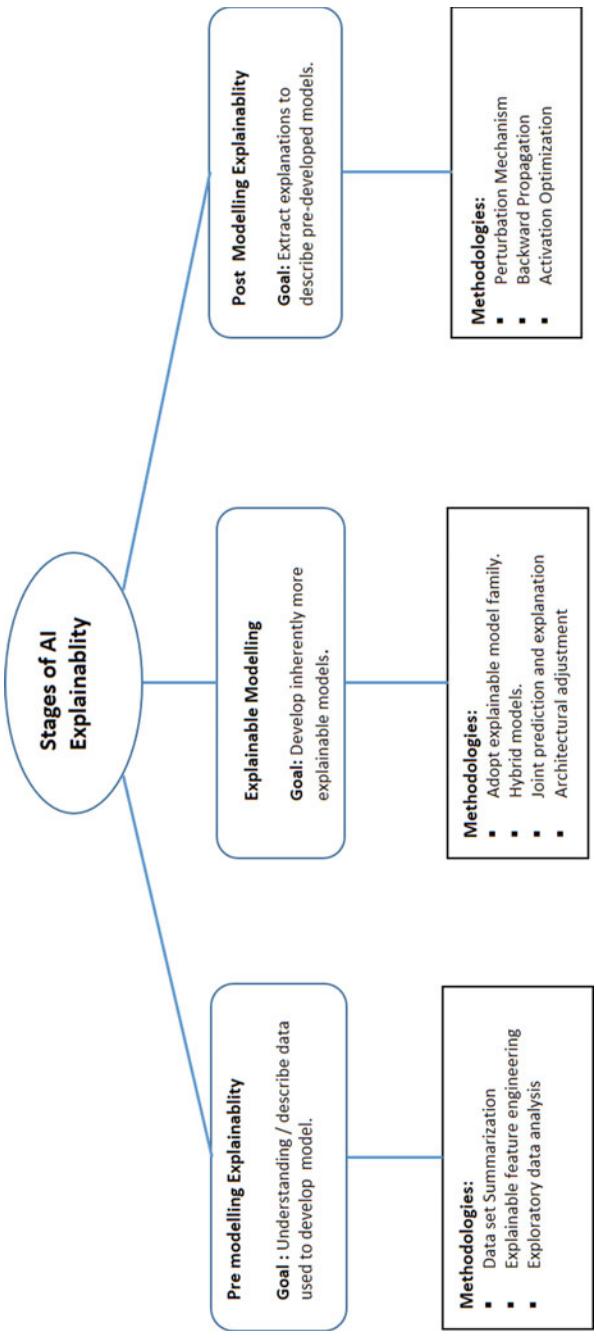


Fig. 4.4 Stages of explainable artificial intelligence (Khaleghi 2022)

datasets are generally complicated and multidimensional, i.e., they contain a significant number of features.

- **Dataset description standardization**: In most cases, datasets are released without adequate documentation. Standardization might help alleviate concerns like systematic bias in AI models and data exploitation by ensuring effective communication between dataset creators and users.
- **Explainable feature engineering**: Feature selection is a technique for limiting the input variable to a model by using only relevant data and removing noise. It is the technique of selecting suitable characteristics for a machine learning model automatically based on the type of challenge. Domain-specific and model-based feature engineering are the two basic ways to achieve explainable feature engineering. To extract and/or identify features, domain-specific techniques rely on domain expert knowledge and ideas obtained from exploratory data analysis. Model-based feature engineering, on the other hand, employs a variety of mathematical models to identify a dataset's underlying structure.

4.3.2.2 Explainable Modeling Methodology

Explainable modeling is also known as In-model Explainability. This methodology is being applied during the explaination of the model. Models which are transparent in nature provide some kind of interpretability themselves. These models are approached based on the domain where they become interpretable, like, algorithmic transparency and decomposability. A model becomes transparent if it can be easily understandable. Some of the transparent models are shown in this section.

- **Logistic and Linear Regression**: It is a type of classification model which is used for predicting a binary variable. However, linear regression would be analog when the dependent variable becomes continuous. Such kind of model would create dependencies between the predicted variables and predictors thus restricting the flexible fitting of the data (Rao 2003). The stiffness exhibited by the model leads to the model being maintained under the category of transparent methods. Although logistic and linear regression entirely meets the properties of transparent models, these methods might also need explainability techniques (including visualization techniques), especially when a model needs to be explained to non-expert audiences (Hellevik 2009). This kind of model are widely used for a long time, which has motivated researchers to create ways for explaining the model predictions to non-expert users.
- **Decision Trees**: This is another kind of model that could easily fulfill all the restrains of transparency. This is a hierarchical structure for making decisions applied to regression and classification problems. It is a simulatable model, whose properties can either make them algorithmically transparent or decomposable. Decision trees are always been categorized under the different types of transparent models (Schetinin et al. 2007). The applications of these models have been closely related to decision-making contexts. This provides an answer to why their complexity and understandability have been considered a crucial matter (Tan et al.

2020) always. A simulatable decision tree is a type of decision tree that can be managed by a human user which indicates that the model size is relatively small and the features and their corresponding meaning are easily understandable. Any increment in model size converts the model into a decomposable one because its size restricts its full simulation. Moreover, a further increase in this size by using the complicated feature is supposed to inject more algorithmic transparency into the model by making it lose the previous characteristics.

- **K-Nearest Neighbor**: The predictions made by the k-nearest neighbor are explained by the data points of the k-neighbor which are neighborhood points whose features were averaged for making the prediction (Peterson 2009). Visualizing the individual cluster which contains alike instances provides an interpretation of why an instance belongs to a particular group. Our analysis suggests that because of the lack of explainability method this is actual and direct at the same time (i.e., it does not produce any illusion of explainability by model approximation) and utilizes the potential of the explainability method in different applications. Some of the recent works have introduced external knowledge and imparted that into the model for improving the interpretation. These XAI methods can fill the gap by introducing the domain knowledge in the model in a model agnostic and transparent procedure. As we mentioned before, it should be kept in mind that the k-nearest neighbor's(KNN) class for providing transparency is dependent on the features, the numbers of neighbors available, and the distance function that measures the similarity in between instances of the data. A high K value hinders full simulation of the performance of the AI model by human users. Similarly, using complex features and distance functions can restrict model decomposability, thus limiting the interpretability only to the transparency of its algorithmic functioning.
- **Bayesian Model**: A Bayesian model is a probabilistic graphical model where its links represented condition-based dependencies in between the set of variables. Considering an example, the Bayesian network can show the relationships between the diseases and their symptoms. If the symptoms are being given, the network could be used for computing probabilities of the existence of different kinds of diseases. Similar to the General Algebraic Modeling System, these kinds of models also provide a clean representation of the relationship between the features and targets. Again, the Bayesian models lack behind concerning the working of transparent models.

4.3.2.3 Post-model Explainability Methodology for AI Models

There arise some cases where the models are not able to meet any of the criteria which makes them transparent in nature, thus a different method needs to be conceived and applied to the model for explaining its decision. Post-model explainability is also known as post-hoc model explainability. The purpose of the post-hoc modeling explainability method is to communicate information that is understandable by the human user and it shows how a model, which is already developed, predicts the output based on provided input. Different kinds of approaches for Post-hoc explaination are shown in Table 4.2.

Table 4.2 Popular approaches for Post-hoc explaination

Types of explainations	Methods	Properties
Local explaination	Rule-based	Here the common sense knowledge and the domain-specific expertise is being represented in the form of plausible rules
	Feature importance	The explainer provides each feature with an important value which shows how much the particular feature is important for the prediction
	Saliency map	It is used in image and video processing and it shows which part is important for the AI decision
	Counterfactual	The explaination could provide a link between that what could have happened when the input of the model is being changed in a particular way
Global explaination	Representation based	It determines the model's reliance on concepts that are semantically logical for humans
	Collection of local explaination	It picks a subset of local explainations to constitute the global explaination after providing a local explaination for every instance of data
	Model distillation	It helps in learning feature shapes which provide the relation between the input feature and the prediction of the model

4.3.3 Types of Post-model Explaination Methods

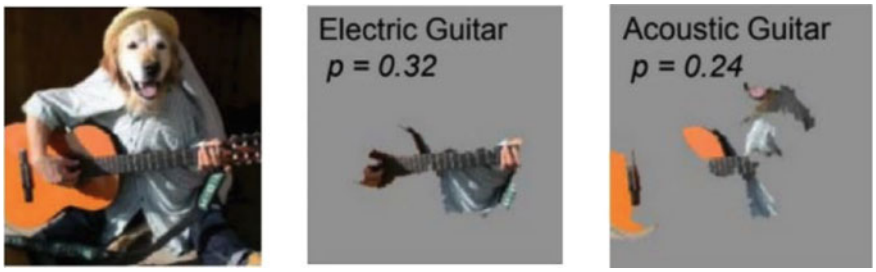
The identified trends surrounding the post-hoc explainability method for different kinds of AI models are as follows.

4.3.3.1 Model-Agnostic Techniques

These type of techniques are designed for plugging it into any kind of model intending to extract some of the information from the prediction process. Some of the techniques

for simplification are being used for generating proxies that mimic their precursor with the motivation of finding out something tractable and less complex. During other times, the main focus is to extract knowledge directly from the model and by directly visualizing it for the betterment of interpretation of their behavior. By following this taxonomy, model-agnostic techniques depend on feature relevance explaination, simplification of the model, and different visualization techniques:

- **SHAP**: SHAP (SHapley Additive exPlainations) is a game theory-based model, which is used for explaining the predictions of any ML model. Using the classic Shapley values from game theory and their related extensions, this method connects optimal credit allocation with locally generated explainations. DeepSHAP (Lundberg and Lee 2017) can be described as a high-speed approximation algorithm for SHAP values in deep learning models that are developed based upon a connection with DeepLIFT. SHAP assigns an equal importance value to every feature for a specific model prediction. The novel components of this method comprise: (1) the identification of an entirely new class of additive feature importance measures, and (2) theoretical results which show that there exists a distinctive solution within this kind of class having a set of desirable properties.
- **LIME**: The LIME framework (Ribeiro et al. 2016) proposed a much more simple approach to all of the previously discussed methods. In essence, LIME perturbs the input data and analyzes the change in the model decisions. For the image classification task, LIME divides the input image into interpretable components after generating a collection of perturbed instances. Next, every perturbed instance is run through the model to get a probability score. Following this step, a simple locally weighted linear model learns on this dataset. Figure 4.5 illustrated the explaination produced by using the LIME method for providing explaination of *Electric guitar* and *Acoustic guitar*. Finally, LIME calculates the super-pixels with the highest positive weights for explaination.
- **Layer-wise relevance propagation**: It is used for visualizing the decision of a convolutional neural network. This method provides a heatmap in the input space which indicates the importance of every pixel that contributes to the final classification. In contrast to the susceptibility maps produced by guided back-propagation,



(Original) (Explaining *Electric guitar*) (Explaining *Acoustic guitar*)

Fig. 4.5 Explaination by using LIME method on images (Ribeiro et al. 2016)

this method can directly highlight the positive contribution concerning network classification of the input space. The layer-wise propagation used by this method is subjected to conservation properties, where the information obtained by a neuron should be re-distributed to the lower layers in a uniform amount. This kind of behavior is also shared by other works on explainations (Landecker et al. 2013; Schütt et al. 2017). Assuming l and m be neurons present in two successive layers, propagating relevance scores R^{lm} in a layer into the neurons of the lower layer is attained by applying :

$$R^{lm} = \sum_{m} \frac{z^{lm}}{\sum_{l}^{z^{jm}}} \tag{4.1}$$

z^{lm} represents the extent to which a neuron l contributed to making the neuron k relevant. The denominator helps in enforcing the conservation property (Bach et al. 2015). The procedure for propagation gets terminated once the input features are covered.

The Post-hoc local explainations and feature relevance methods have gradually become the most adopted method for explaining deep neural networks.

- **DeepLIFT**: DeepLIFT (Deep Learning Important FeaTures) is a method that is used for decomposing the network predictions on a particular input by the back-propagation of the contributions of all neurons to every input feature (Li et al. 2021). In essence, DeepLIFT compares the activation of every neuron with its *reference activation* and allocates the contribution scores based on the differences. By having an option to separately consider the positive and negative contributions, DeepLIFT can reveal the dependencies missed by other approaches. The scores are efficiently computed in a single backward pass.

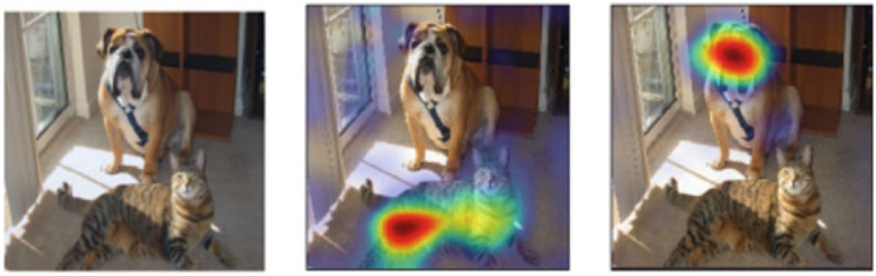
Some of the Explainability methods for visualization of the AI prediction are shown in Table 4.3.

4.3.3.2 Model-Specific Techniques

Model-specific techniques are based on the details of the applied machine learning or deep learning model's distinctive structures. These techniques are employed specifically for a model architecture, such as a convolutional network model CNN. These methods take advantage of the internals of machine learning models such as neural networks and apply a reverse engineering approach to explain how the relevant DL or ML algorithm is making the decision. The benefits of employing model-specific models are that they allow us to gain a better understanding of the choice by revealing the model's internal workings, as well as allowing us to create a more tailored explainable model. On the other hand, such models necessitate going through the entire structure of the model, putting the model's performance at risk because the ML or DL model will be recreated. The structure of the algorithms is usually exam-

Table 4.3 Visualization methods for explainable Artificial Intelligence

Methods	Description	Uses	Limitations
Deep SHAP (Lundberg and Lee 2017)	It is used for computing the SHAP values based on game theory. The algorithm is fast and is connected with DeepLIFT using multiple background samples	It helps in determining attributions for non-neural models like Decision trees, Support Vector Machines(SVM), etc	Mathematically articulated problems arise when Shapley values are interpreted as feature importance measures
Deep LIFT (Gilpin et al. 2018)	Here a reference input is used for computing the reference value of the hidden units. It avoids placing potentially misleading importance on bias terms. It consists of two variants—Rescale rule and RevealCancel. RevealCancel treats positive and negative contributions in a neuron	Rescale is sometimes related to e-LRP but it can not be applied to models that involve multiplicative rules. RevealCancel manages such situations by using RevealCancel for convolutional and Rescale for fully connected layers which helps in reducing the noise	It can not be applied to models that involves multiplicative rules
Integrated Gradients (Qi et al. 2019)	It aims to explain the relationship between a model's predictions in terms of its features. It works by computing the average gradient because the input differs from the baseline to the value of the actual input, unlike the Gradient input which uses a single derivative at the input.	It has many use cases which include understanding the feature importance and debugging the model performance, it is also used for augmenting the accuracy metrics.	The heatmap generated by the integrated gradients is diffused
Saliency Maps (Simonyan et al. 2013)	These compute the absolute value of the partial derivative of the output neuron w.r.t the input features for detecting the most influential feature. It shows the importance of a pixel to the human visualization system	This method captures the instinct such that the information of a location is proportional to the activation level. It processes the images for differentiating the visual features present in images.	It is noisy because deep neural networks do not filter out irrelevant features during forward propagation
Layer wise relevance propagation (LRP) (Montavon et al. 2019)	The purpose of LRP is to provide an explaination of a neural network's output with context to the input's domain It distributes the prediction score layer with the layer having a backward pass on the network by using specific rules like e-rule when the numerical stability is ensured	It brings explainability to the AI decisions that help to scale up the potential of complex deep neural networks. LRP works by backward propagation of the network, using a set of designed rules for propagation. This method does not interact with training the network, so it could be easily applied to pre-trained classifiers	This method is limited to the CNN models with ReLU activation
Grad-CAM (Selvaraju et al. 2017)	A gradient-based class activation map is produced by using the gradients of the targeted concept as it goes to the last convolution layer	This method applies to CNN which includes the fully connected layers, structured output, and reinforcement learning	Lack of explaining decisions produced by deep networks in the domains like natural language processing, and reinforcement learning



(Original) (Explaining *Cat*) (Explaining *Dog*)

Fig. 4.6 Grad-CAM showing visualization heat map for differentiating the images as *Cat* and *Dog* (Selvaraju et al. 2017)

ined using model-specific approaches. Some of the model specific techniques are illustrated as follows.

- **Grad-CAM**: Grad-CAM is the acronym for Gradient weighted Class Activation Mapping. Several visualization approaches and gradient-based methodologies (Zeiler and Fergus 2014) have been used for providing explainations in deep learning solutions. Gradient weighted Class Activation Mapping applies the gradient of the targeted conception into the final convolutional layer of the model for building a localization map that highlights the main region of the input image. Grad-CAM observes both the forward and backward passes and gives better visualizations by representing a local influence of the input image in predicting the output classes. We only consider backward passes in the de-convolution methods. This algorithm is a class discriminated localization technique that brings out higher resolution visual explaination from the CNNs (Convolutional Neural Networks). A Grad-CAM based visual explaination of the AI model identified as *Cat* and *Dog* using the heatmap is shown in Fig. 4.6.
  For producing the localization map for a specific image class i, Grad-CAM finds out the gradient score g^i before the softmax layer with reference to the feature map f^k :

$$g^i(f^k) = \frac{dy^i}{df^k} \tag{4.2}$$

  here, k is denoted as an index of the channel and thus the averaged gradient score is denoted as m^c where:

$$m^c = \frac{1}{UV} \sum_i \sum_j g^i(f^k) \tag{4.3}$$

  here U and V represent the length and the width of an image provided as input. Last feature map of the class is represented as N^k. Where $N^k \subset P^{U*V}$, and, P is real number. Thus from Equation 2, The Grad-CAM can be represented as Equation

3:

$$Grad - CAM = relu(\sum_{k=1}^{\infty} m^c)N^k \tag{4.4}$$

The weight of the feature maps is being used and the weighted sum is being calculated which generates the final heat map.

- **Integrated Gradient**: Integrated Gradient (IG)) is a method for computing the gradient of the predictions of a model to its input features and does not require us to modify the original deep neural network (Saibi et al. 2006). We can apply IG to any model that is differentiable like images, texts, or any type of structured data. Methods such as Integrated Gradients are model-specific and require knowledge about the internal model for computing the gradients of the model layers.
- **Saliency maps**: The objective of computing saliency maps is to find out the image regions that are different (or conspicuous) from their neighbors based on the image features (Thompson and Bichot 2005). Given an image, we first compute the basic image features such as color, orientation, intensity, etc. These processed images are then used to construct Gaussian pyramids that are in turn used to create feature maps at different scales. Next, the saliency map is created by taking the average of all the feature maps (Zeiler and Fergus 2014). This technique measures information at the end of every network scale that is subsequently combined to form a single saliency map.

4.4 Metrics for Explainable Artificial Intelligence

Different evaluation metrics should accompany the measures adopted in XAI research (Hoffman et al. 2018). The evaluation metrics are expected to achieve the identified properties of explainability as objectives. The quantitative metrics for both model-based and example-based explainations are largely used to assess the ease with which they can be understood, whereas the quantitative metrics for attribution-based explainations are mostly used to assess the fidelity with which they can be explained. We have identified the different properties of explainability by studying the corresponding definitions. These identified properties of explainability are used as objectives that we expect the evaluation metrics to achieve. Standardization organizations such as ISO have also raised the need for evaluation metrics for measuring the reliability of AI systems (Russakovsky et al. 2015). Their standardization document has outlined different challenges related to the implementation and the use of AI systems. Over-reliance and under-reliance on the AI system are the primary concerns raised by this work. Over-reliance occurs when a user is more reliant on the automation aspects without considering the associated limitations. This can have adverse outcomes (Yang et al. 2019). Under-reliance occurs when the user frequently disagrees with the correct AI system decisions.

4.4.1 Evaluation Metrics for Explaining AI Decisions

As discussed in Sect. 4.3, the methods of explaination are broadly classified into three types namely Model-based Explainations, Attribute-based Explainations, and Example-based Explainations. In this section, the quantitative metrics are being discussed for measuring the qualities of these explaination methods. Apart from the evaluation of the AI explaination methods, proper selection of the evaluation metrics plays a major role in evaluating the system accurately. Different types of metrics that evaluate the extent of the explainability by different methods are as follows:

- **Subjective metrics**: It is designed for questioning the users based on tasks and the explainations provided, these questions are asked when the task gets executed or afterward for obtaining the subjective response from the user on the explainations. Some of the examples of these types of metrics are the confidence, and trust of users that have an enormous grasp over the focal points for evaluating the explainable system. Hoffman et al. (2018) proposed a metric that is used for the subjective evaluation of an AI system. It considers factors like user trust, understanding, and satisfaction. Zhou et al. (2016) have looked over the factors like the uncertainty that affect the trust of users in informed machine Learning decision makings. They established that explaination generated because of influence in the training data points remarkably affects the user's trust in case of informed decision making.
- **Objective metrics**: It mentions the objective information of a task to a user before or after the task is being performed. Such kinds of examples are human metrics, which include behavior and physiological measures of humans when informed decision making takes place, another such metric is task-related metrics, which include time length for completing the task and performance of the task. Schmidt and Biessmann (2019) showed that fast and accurate decisions mean instinctively understanding the explainations provided. It resulted in deriving a trust metric that is based on the explainability metrics.
- **Computational metrics**: These metrics are known as mathematical indicators for determining the quality of explainations generated by an XAI system. The measurement of these kinds of explainations is generally being carried out by using necessarily developed equations. Thus, these metrics may be used without any kind of human intervention as guidelines for preparing the explaination techniques.
- **Cognitive metrics**: The explainations provided to the end-users are being measured by using cognitive metrics. The assessment of human subjects is a blunt indication of explainations, as we know the initial goal of XAI is to convey the reasons behind machine judgments to people.

Accuracy is one of the most commonly used metric (Rosenfeld 2021), it is very easy for understanding although noticing only these metrics would give an incomplete suggestion regarding the performance of a model. Multiple established metrics are there which would provide a thorough insight into the performance of the model. The metrics used for quantifying the explainations are generally very specific to the different types of machine learning problems and models. Some of the widely used

Table 4.4 Metrics for quantitatively explaining AI decisions

Types of explaination	Metrics	Explaination properties
Model-based explainations	Model size (Guidotti et al. 2018)	Simplicity
	Interaction strength (Markus et al. 2021)	Simplicity
	Level of agreement (Lakkaraju et al. 2017)	Clarity
Attribution-based explainations	Effective complexity (Nguyen and Martínez 2020)	Broadness and simplicity
	Recall of important features (Ribeiro et al. 2016)	Soundness
	Selectivity and continuity (Montavon et al. 2018)	Soundness and clarity
	Mutual information (Nguyen and Martínez 2020)	Broadness and soundness
	Sensitivity (Montavon et al. 2018)	Soundness
Example-based explainations	Diversity (Nguyen and Martínez 2020)	Simplicity
	Non-representatives (Nguyen and Martínez 2020)	Simplicity and completeness

metrics are Loss, Confusion Matrix, Accuracy, Mean Absolute Error, Root Mean Square Error, Accuracy, etc.

Table 4.4 shows the metrics for quantitatively explaining the AI decisions for explaining the classification.

4.5 Use-Case: Explaining Deep Learning Models Using Grad-CAM

As a use-case of XAI, we implemented some of the visual methods of XAI for explaining the learned models for classifications of fresh tea leaves (Banerjee et al. 2022). We have applied a well-known technique for introducing visual explainations of predictions from Neural Network-based models, which makes them more interpretable. We have used transfer learning for classifying different types of tea leaves. The aim here is to reveal the underlying features that are responsible to explain the relationship between the predictions of a tea-leaf classification model by using popular visualization techniques. In this work, the Grad-CAM method has been used that use the gradients of the targeted concept. We performed our experiment on our image dataset created by imaging and labeling the freshly harvested tea leaves

from *Banuri experimental tea farm*, Palampur, India. Our data consists of 965 tea leaf images where the combination of 1-Leaf 1-bud consists of 261 data samples, 2-Leaf 1-bud has 174 data samples, 3-Leaf 1-bud have 279 data samples, 4-Leaf 1-bud has 199 data samples, and 5-Leaf 1-bud has 52 data samples. Because of the class imbalance in our dataset, we have implemented data augmentation, which reduced the class imbalance. We used the final layer of convolution for producing a localization map that highlights the most important regions of the image while determining the class of the tea-leaf. While performing our experiments, we found that the Grad-CAM method takes out features from the pre-trained VGG16 model. We implemented transfer learning by using well-known pre-trained backbone models like VGG16 and InceptionV3. For our tea leaf dataset, we discovered that the performance evaluation metrics of our built model on 1L1B(1-Leaf 1-bud) have overall best performance, with precision, recall, and F1-scores of 0.90, 0.90, 0.90. Our 2L1B data (2-Leaf 1-bud) has precision, recall, and F1scores of 0.73, 0.62, and 0.67, respectively. Whereas 3L1B (3-Leaf 1-bud) has a precision of 0.80, recall of 0.82, and F1-score of 1. Moreover, 4L1B (4-Leaf 1-bud) has a comparatively low precision of 0.69, but has a good recall of 0.92, and has an F1score of 0.79. 5L1B (5-Leaf 1-bud) has a precision of 1 and has a comparatively poor recall of 0.18, and an F1score of 0.30 which shows the model has certain limitations in the identification of the actual distinguishing features responsible for classifications for tea leaves. Our trained classification model, which employs VGG16, predicts an average of 0.83, 0.69, 0.74 precision, recall, and F1-score, with an accuracy of 80%. And by using InceptionV3 our model could able to have achieved an accuracy of 76.41% respectively. We obtained explainations based on the classification in our dataset of the tea leaf images by providing the pre-trained models as input to the Grad-CAM pipeline for producing class-specific heat maps. We used Grad-CAM as an explaination method for explaining our prediction of the model. The model generates explainations for different leaf image categories. Grad-CAM exclusively highlights the important portions of the image which are of utmost importance for predicting the class of the tea leaf image. Figure 4.7 shows the generated heatmaps for explaining the most important image regions which are influential in predicting the result using pre-trained model VGG16 and InceptionV3 models.

The heatmap produced by Grad-CAM for the pre-trained VGG16 model focuses particularly on the Leaf and Bud portion of the input image, whereas the heatmap generated by Grad-CAM with pre-trained InceptionV3 model has an inappropriate focus on the stem region of the leaf image. Grad-CAM uses the final feature map of the model for creating heatmaps that highlight the image pixels responsible for the prediction of the image class.

4.6 Challenges and Future Directions

Explainable AI is an active area of research for handling the black-box nature of deep learning models. The most recent literature survey, reveals that though there

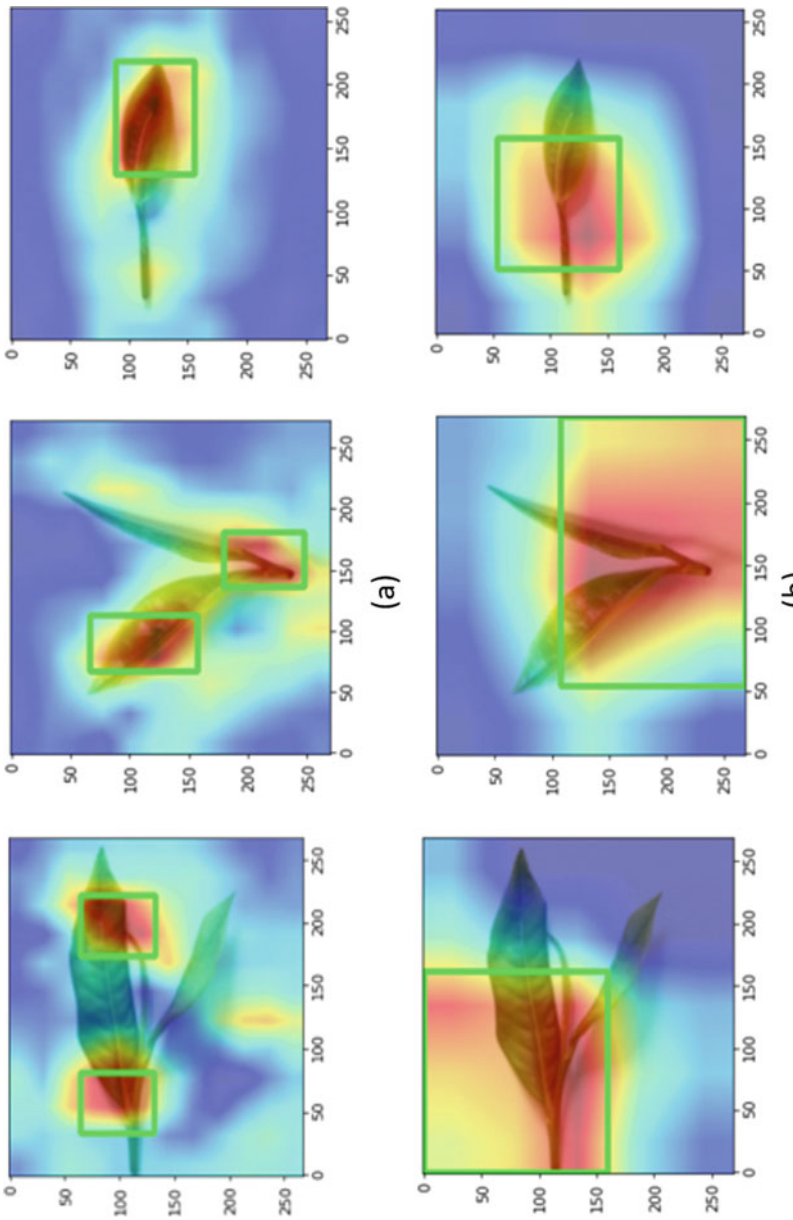


Fig. 4.7 Explaining the prediction by using Grad-CAM and pre-trained **a** VGG16, and **b** InceptionV3 models (Banerjee et al. 2022)

are several works published in this domain of research there are still many important problems to be resolved by the research communities. Though there are a lot of AI black-box methods are there which need explainations but due to high-dimensional issues, the image classification and object detection problem in computer vision is challenging. The black-box models that yield highly accurate results normally need to be explained using visual explaination for better perception of the underlying working of the model behind the same. However, the main challenge for visual explainations of a classifier's output arises from the complex nature of the AI model and underlying data. Since the visual explaination method relies on the algorithm for the generation of human perceptible feedback which becomes subjective in nature. Moreover, it is very challenging to properly explain complex classification models accurately. The classifiers which provide good performances are gradually becoming more complex due to the involvement of the numerous parameters and the operations performed by them which in turn makes them complex and tough to explain. Because of the different architectures of the AI methods, the problem to design an effective framework for explaining the underlying decision process increases. Other challenges include the use of multiple types of data for training the models. For explaining an AI model, the most common strategy is to trace back to the input data. Different types of data require different types of explainations. Though the AI model based on image data is relatively easier for generating visual interpretations of the decision process, the same using the textual, speech, or nominal data is difficult to get explained similarly. Even in image data-based AI models, there are no objective metrics available that measures the quality of explainations. This paves the way for further research in this area. In this section, we endeavored to identify a few main research questions in the field of XAI. These questions include some necessary aspects like how do we evaluate the Extent of Explaination for an AI model? While referring to the AI perspective, metrics are defined as any quantification of the extent of explaination which helps in evaluating its quality and suitability.

Evaluation metrics explain the model's performance. It is used in measuring the quality of the explaination. Though there are multiple metrics used for tasks of classification, ranking, clustering, regression, modeling topics, etc. there are very limited metrics are there for measuring the extent of explaination in different data types and AI tasks. This in turn forms another research question whether a method of explaination can be devised that is invariably applicable to any data type or AI model?

There are different methods of explainable AI that target explaining the decision-making of the AI system and thus can identify the problems associated with the underlying model. But it needs to be investigated how explainable AI models can help in improving the prediction accuracy or inhibiting the failure points in real-life problems. XAI consists of a set of frameworks and tools which helps in interpreting and understanding the predicted output of the AI models. It is also imperative to study the applicability of the explaination method on a local instance of data or globally on an entire dataset. There are techniques where an AI model can be explained locally and globally based on the given input data but rarely any specific XAI method is available which can effectively evaluate the quality of explainations both in the local

and global dataset. Also, there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these kinds of methods for increasing trust in them. Moreover, these kinds of metrics could be used for standardizing the XAI solutions.

4.7 Conclusion

Explainable AI (XAI) is an emerging technological paradigm of which most enterprises are conscious. The methods and processes of XAI provide several advantages. Explainability in pre-modeling is a feasible but under-focused approach for avoiding transparency problems. Pre-modeling explainability methods mainly focus on the explainability of data rather than the model itself. Whereas Post-modeling/Post-hoc explainability is a collection of different types of methods with a common goal of gaining a better understanding of the working of the trained model. Based on Post-hoc explaination the methods are classified into model-agnostic and model-specific techniques. Moreover, the metrics used for explaining the AI decisions are quantitatively evaluated as Subjective metrics, Objective metrics, Computational metrics, and Cognitive metrics for evaluating the AI system accurately. Although there is a necessity for more quantitative evaluation metrics which would provide a comparison between different types of explainability methods and would quantify the acceptance of these methods for enhancing trust in them. For increasing transparency in the developed model, it is necessary to produce an intuitive explaination. In this chapter, we have presented a comprehensive overview of the methods and metrics for explaining decisions made by AI models. We also covered the taxonomy of XAI in ample detail and discussed different strategies used for providing explainations behind the working of the data-based learned models. We have presented a selected overview of works to assist researchers and practitioners in understanding insights, accessible resources, and unresolved difficulties in using XAI methodologies. A use-case of implementing a popular XAI visualization method is also been demonstrated in this chapter. Though the research work in the area of XAI is in full swing but still has many gray areas to be addressed by the global AI communities. The research directions section in this chapter is an endeavor to summarize the identified research gaps and unanswered research questions for prospective XAI researchers.

References

Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)

Antwarg, L., Miller, R.M., Shapira, B., Rokach, L.: Explaining anomalies detected by autoencoders using shap. arXiv preprint arXiv:1903.02407 (2019)

Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:1909.03012 (2019)

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130, 140 (2015)

Banerjee, P., Banerjee, S., Barnwal, R.P.: Explaining deep-learning models using gradient-based localization for reliable tea-leaves classifications. In: 2022 IEEE Fourth International Conference on Advances in Electronics, Computers and Communications (ICAECC), pp. 1–6. IEEE (2022)

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U.: Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911 (2017)

Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019)

Christopher Frey, H., Patil, S.R.: Identification and review of sensitivity analysis methods. Risk Anal. **22**(3), 553–578 (2002)

Cortez, P., Embrechts, M.J.: Opening black box data mining models using sensitivity analysis. In: 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 341–348. IEEE (2011)

Cortez, P., Embrechts, M.J.: Using sensitivity analysis and visualization techniques to open black box data mining models. Inf. Sci. **225**, 1–17 (2013)

Dignum, V.: Responsible artificial intelligence: designing AI for human values (2017)

Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. In: 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2020)

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

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Statist. **24**(1), 44–65 (2015)

Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT press (2016)

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 1–42 (2018)

Hellevik, O.: Linear versus logistic regression when the dependent variable is a dichotomy. Qual. Quant. **43**(1), 59–74 (2009)

Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)

Jain, A., Ravula, M., Ghosh, J.: Biased models have biased explanations. arXiv preprint arXiv:2012.10986 (2020)

Khaleghi, B.: The how of explainable AI: explainable modelling. https://towardsdatascience.com/the-how-of-explainable-ai-explainable-modelling-55c8c43d7bed

Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: International conference on machine learning, pp. 1885–1894. PMLR (2017)

Krarup, B., Cashmore, M., Magazzeni, D., Miller, T.: Model-based contrastive explanations for explainable planning (2019)

Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. arXiv preprint arXiv:1707.01154 (2017)

Landecker, W., Thomure, M.D., Bettencourt, L.M., Mitchell, M., Kenyon, G.T., Brumby, S.P.: Interpreting individual classifications of hierarchical networks. In: 2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp. 32–38. IEEE (2013)

Li, J., Zhang, C., Zhou, J.T., Fu, H., Xia, S., Hu, Q.: Deep-lift: deep label-specific feature learning for image annotation. IEEE Trans, Cybern (2021)
Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
Markus, A.F., Kors, J.A., Rijnbeek, P.R.: The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J. Biomed. Inf. **113**, 103,655 (2021)
Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning, pp. 193–209 (2019)
Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Process. **73**, 1–15 (2018)
Next Move Strategy Consulting (NMSC): explainable AI market size , share, forecast, industry analysis report | 2021 - 2030. https://www.nextmsc.com/report/explainable-ai-market
Nguyen, A.P., Martínez, M.R.: On quantitative aspects of model interpretability. arXiv preprint arXiv:2007.07584 (2020)
Peterson, L.E.: K-nearest neighbor. Scholarpedia **4**(2), 1883 (2009)
Qi, Z., Khorram, S., Li, F.: Visualizing deep networks by optimizing with integrated gradients. In: CVPR Workshops, vol. 2 (2019)
Rao, S.J.: Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis (2003)
Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144 (2016)
Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386 (2016)
Rosenfeld, A.: Better metrics for evaluating explainable artificial intelligence. In: Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems, pp. 45–50 (2021)
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
Saibi, H., Nishijima, J., Ehara, S., Aboud, E.: Integrated gradient interpretation techniques for 2D and 3D gravity data interpretation. Earth Planets Space **58**(7), 815–821 (2006)
Samek, W., Müller, K.R.: Towards explainable artificial intelligence. In: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 5–22. Springer (2019)
Schetinin, V., Fieldsend, J.E., Partridge, D., Coats, T.J., Krzanowski, W.J., Everson, R.M., Bailey, T.C., Hernandez, A.: Confident interpretation of Bayesian decision tree ensembles for clinical applications. IEEE Trans. Inf. Technol. Biomed. **11**(3), 312–319 (2007)
Schmidt, P., Biessmann, F.: Quantifying interpretability and trust in machine learning systems. arXiv preprint arXiv:1901.08558 (2019)
Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nat. Commun. **8**(1), 1–8 (2017)
Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626 (2017)
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
Suman, R.R., Mall, R., Sukumaran, S., Satpathy, M.: Extracting state models for black-box software components. J. Object Technol. **9**(3), 79–103 (2010)

Tan, S., Soloviev, M., Hooker, G., Wells, M.T.: Tree space prototypes: another look at making tree ensembles interpretable. In: Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, pp. 23–34 (2020)

Thompson, K.G., Bichot, N.P.: A visual salience map in the primate frontal eye field. Prog. Brain Res. **147**, 249–262 (2005)

Tukey, J.W., et al.: Exploratory Data Analysis, vol. 2. Reading, MA (1977)

Van Lent, M., Fisher, W., Mancuso, M.: An explainable artificial intelligence system for small-unit tactical behavior. In: Proceedings of the National Conference on Artificial Intelligence, pp. 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999 (2004)

van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103,404 (2021)

Yang, F., Du, M., Hu, X.: Evaluating explanation without ground truth in interpretable machine learning. arXiv preprint arXiv:1907.06831 (2019)

Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision, pp. 818–833. Springer (2014)

Zhou, J., Arshad, S.Z., Yu, K., Chen, F.: Correlation for user confidence in predictive decision making. In: Proceedings of the 28th Australian Conference on Computer-Human Interaction, pp. 252–256 (2016)