Chapter 5
Evaluation Measures and Applications for Explainable AI

Mayank Chopra and Ajay Kumar

Abstract Machine learning advances, particularly deep learning, have enabled us to design models that excel at increasingly complicated tasks. Because of the growing size and complexity of these models, it's becoming more difficult to grasp how they arrive at their forecasts and when they go incorrect or even worse. Now, think of a situation in which we humans could open these black-box learning models and translate the content into a human-understandable format. This is known as Explainable Artificial Intelligence and there has been a lot of research in this field over the last few years mainly focusing on how to explain different types of models. The advancement of this research, raised a very important query: "Why does a model need to be explained?" So, the most accurate answer to this question is "TRUST". TRUST that the models are making the correct decisions over the correct assumptions. TRUST that we can tell what happened when a model fails. TRUST that we can do on a model implemented on a large scale that the predictions are made in line with expectations. It's hard to trust a system that's not transparent about its internal processes. This paper discusses the evaluation measures and application areas of XAI. Some XAI-related concepts were also mentioned.

Keywords Explainable artificial intelligence · Interpretable machine learning · Machine learning · XAI · Deep learning

5.1 Introduction

It is crucial for an artificial intelligence (AI) program to be able to give accessible justifications that properly justify its conclusions in circumstances when people are expected to make key decisions powered by AI. An adequate justification can

M. Chopra (✉) · A. Kumar
Department of Computer Science and Informatics, Central University of Himachal Pradesh (H.P.), Dharamshala, India
e-mail: mayankchopra.it@gmail.com

A. Kumar
e-mail: ajaykr.bhu@hpcu.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_5

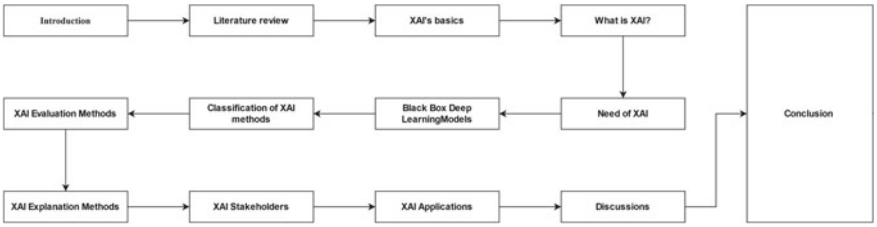

Fig. 5.1 Work Flow of the paper

enhance the system's trust, enabling better human-AI collaboration (Villata et al. 2013). Explanations can assist people in determining how much, they must believe the explanation source. In both industry and academia, AI algorithms are gaining a lot of traction (Attaran and Deb 2018). Since Machine Learning (ML)-based procedures are swamped with thousands of hardly analyzable variables to be improved during the training phase, a variety of these procedures are frequently referred to as "black-box" algorithms (Stepin et al. 2021). Because of this, the algorithm's outcome is difficult to explain. The failure to comprehend these automated judgments diminishes users' trust in such systems, reducing their usability (Ribeiro et al. 2016). Furthermore, many current explainable AI systems deliver summaries of automatically generated forecasts instead of complete explanations (Rudin 2019). As a consequence, the necessity to justify automated judgments with a good description of why the algorithm makes a specific decision has fueled rapid growth in the explainable AI research field (Anjomshoae et al. 2019).

In this paper, we have briefly reviewed the Explainable Artificial Intelligence Literature followed by the XAI's Evaluation Methods which is followed by the XAI's Explanation Methods. After that, we revisited the stakeholders and the applications. Some proposed frameworks in the field were also taken into consideration and in the end, we have concluded our study. Figure 5.1 shows the workflow of our paper.

5.2 Literature Review

The ideas of understandability and comprehensibility are difficult to describe precisely; however, numerous attempts have been made, with the most famous works comprising (Lipton 2018; Doshi-Velez and Kim 1702). Gilpin et al.'s (Gilpin et al. 2018) work is the other effort to identify the key concepts of understandability in machine learning. The researchers created a taxonomy for defining accuracy techniques for neural nets into three classifications while focusing primarily on deep learning. Between 2004 and 2018, Adadi and Berrada did a thorough analysis by collecting and processing 381 separate academic papers. They categorized all research on explainable AI into four categories, outlining the necessity for additional approval in the XAI domain including more human–machine interaction. After

noticing the community's habit of researching explainability only in the aspects of modeling, they called for incorporating it into other aspects of machine learning (Kumar and Chatterjee 2016). Finally, they suggested a potential area of study including the fusion of existing explainability approaches (Adadi and Berrada 2018). Murdoch et al. (Murdoch et al. 2019) published a fact sheet following the discovery of the shortage of regularity and a way to measure the effectiveness of classification strategies. They devised an interpretability paradigm to overcome the previously indicated gap. Classification power, describing accurateness, and appropriateness are three types of metrics presented by the Relevant, Descriptive, and Predictive framework for evaluating classification methodologies (Kumar et al. 2021). According to recent research, a new type of arrangement was provided that first differentiated truthful and subsequent methods, and then generated sub-classes (Arrieta et al. 2020). A separate ontology was created specifically for deep learning classification techniques due to the large number of them (Gautam and Chatterjee 2021).

5.3 Basics Related to XAI

5.3.1 Understanding

Understanding is linked to the human ability to spot connections and the perspective of a dilemma, and it is a prerequisite for explanations. Mechanistic understanding ("How does anything operate?") and functional understanding ("What is its objective?") are two types of understanding.

5.3.2 Explicability

Explicability refers to the ability to examine the characteristics of an AI system.

5.3.3 Explainability

Explainability goes beyond explicability by aiming to make the perspective of an AI system's logic, model, or findings for a judgement outcome available, so that beings can comprehend it.

5.3.4 Transparency

An AI system is transparent if its algorithmic activity in terms of decision outcomes or procedures can be comprehended by a human.

5.3.5 Explaining

Using explicability or explainability to enable a human to recognize a model and its intent is referred to as explaining.

5.3.6 Interpretability

Interpretability implies how an AI system's judgement could be clarified globally or locally and that the system's intent could be comprehended by a human.

5.3.7 Correctability

Correctability denotes the capacity of an AI system to be focused by a human actor to guarantee the right choices.

5.3.8 Interactivity

Interactivity occurs when it is feasible to progressively investigate and accommodate the inner structure of a model (correctability). This differs from local and global understandability, which relates to the presentation of outcomes and routes.

5.3.9 Comprehensibility

Comprehensibility, like interpretability, is based on local and global interpretations as well as functional knowledge. Furthermore, understandable AI satisfies interactivity. Both interpretable demonstration and invasion are viewed as essential elements for in-depth comprehension and thus as prerequisites to comprehensibility.

5.3.9.1 Human-Artificial Intelligence System

A human-AI system includes both computational elements and a human user who must come together to accomplish a purpose.

5.4 What is Explainable AI?

Explainability is a concept that stands at the crossroads of numerous fields of active AI research, with an emphasis on the following domains.

5.4.1 Fairness

Can we verify that choices taken by an AI system were done consistently?

5.4.2 Causality

Can one learn a system from facts that not only makes the right predictions but also offers some understanding of the core events?

5.4.3 Safety

Can we have confidence in the reliability of our AI system without recognizing how it makes its presumptions?

5.4.4 Bias

How can we be confident that the AI system hasn't picked up a distorted view of the world due to flaws in the training data or objective function?

5.4.5 Transparency

Everyone has a right to be informed about changes that influence us in ways, formats, and languages that we comprehend.

An XAI, also known as a "Transparent AI" or "Interpretable AI", is an AI whose activities are simple for humans to comprehend and evaluate. A civic privilege to explain can be implemented using XAI.

5.5 Need for Transparency and Trust in AI

The black box AI systems have found their way into many of today's modern implementations. Transparency and explainability are not critical requirements for machine learning models used as long as the overall efficiency of these systems is adequate. Even if these systems fail, the implications are unexceptional. As a result, the necessities for trust and openness in these types of AI systems are relatively low. The scenario is different in safety–critical applications. In this case, the opaqueness of ML techniques may be a restricting or indeed rejecting component. Particularly when a single misjudgment can endanger human life and health or lead to significant revenue damages, depending on an information system with unintelligible logic will not be an alternative. This lack of transparency is among the causes why the application of machine learning to areas such as healthcare is extremely careful than its application in the consumer, electronic commerce, or media industries.

5.6 The Black Box Deep Learning Models

The method of developing interpretations for AI system behavior will vary based on the type of ML techniques used: techniques that produce implicitly decipherable models vs deep learning algorithms that are intricate information and understanding methods and produce models that are implicitly indecipherable to actual users.

ML techniques such as Bayesian classifiers, decision trees, spare linear models, and additive models produce decipherable models in the sense that model components can indeed be instantly examined to comprehend the model's inferences. These technique makes use of relatively small internals, and also provide visibility and traceability in their decision-making. As long as the model is precise for the classification process, these strategies offer awareness of the AI system's decision-making.

Deep learning algorithms, one on either side, are a class of machine learning technique that sacrifices clarity and interpretability for the predictability. These techniques are now used to create applications such as consumer behavioral forecasting associated with high inputs, voice recognition, natural language processing, and computer vision.

The lack of transparency and understandability in the Deep Learning Algorithms makes them a black box. The black box model is a model which performs its predictions on its own without explaining anything for humans to understand.

The Black Box Problem occurs when artificially intelligent processor architectures are vague.

This figurative language is based on the notion that the function of a system may be explained by "gazing within." Although, modern computing systems are composed of well-known hardware components that present no physical barriers to peering inside. However, they may be seen as mysterious in the sense that it is tough to comprehend how such devices are designed.

With time prediction techniques have made considerable latest improvements in tackling the transfer among analysis and prediction affiliated with deep learning models—these techniques estimate deep-learning black-box models with simplified decipherable models that could be examined to clarify the black-box models.

These techniques are known as XAI because they transform black-box models into crystal models. They are gaining popularity because they allow Ai systems to undertake both predictive performance and interpretability goals.

5.7 Classification of XAI Methods

In the literature, several categorizations have been suggested to categorize various explainability techniques. In general, categorization methods are not exquisite; they can vary greatly based on the methodological features and can be categorized into several overlapping or non-overlapping classes at the same time. Various classification methods and taxonomies are briefly mentioned here, and a flow diagram for them is exhibited in Fig. 5.2 Flow diagram of the classification of XAI Methods.

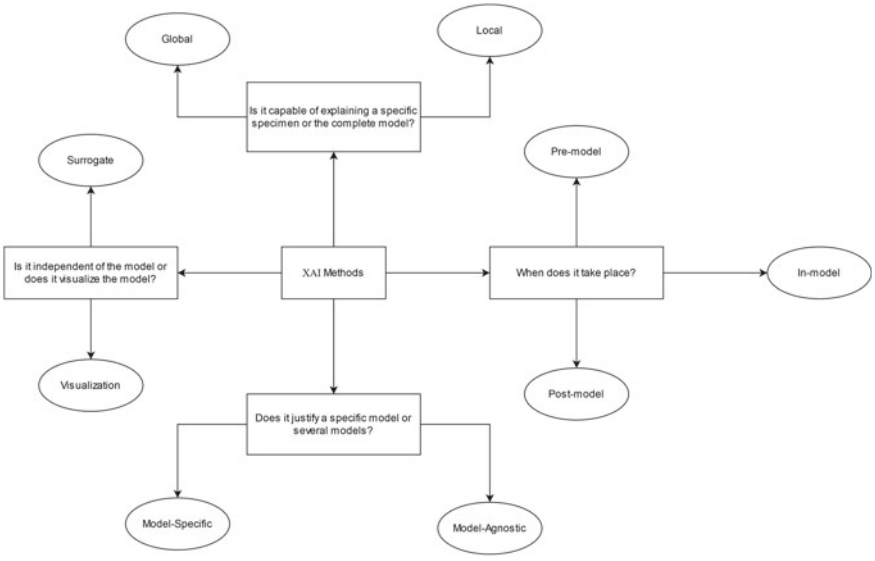

Fig. 5.2 Flow diagram of the classification of XAI Methods

5.7.1 Global Methods Versus Local Methods

Local explainable techniques are only useful for single model output. This can be accomplished by developing strategies for explaining a specific estimation or conclusion. Global methods, on the other hand, focus on the interior of a prototype by leveraging the knowledge base about the prototype, the mentoring, and the related data. It attempts to clarify the model's conduct in particular.

5.7.2 Surrogate Methods Versus Visualization Methods

Surrogate methods are ensembles of various models that are used to examine other black-box models. The black box models can be effectively comprehended by analyzing the surrogate model's judgments and correlating the black-box model's and surrogate model's judgments. Surrogate techniques are illustrated by the decision tree. The visualization techniques are not just a separate model, but they help clarify some aspects of the models through the image process, such as activation maps.

5.7.3 Model Specific Versus Model Agnostic

The metrics of the specific models are used to determine model-specific analysis techniques. The GNNExplainer (graph neural network explainer) is a type of model-particular explainability in which the complex nature of data depiction necessitates the use of the GNN specifically. Model Agnostic techniques are usually used in afterward assessment and are not restricted to a specific model architecture. These techniques lack explicit direct exposure to underlying model weights and architectural metrics.

5.7.4 Pre-Model Versus In-Model Versus Post-Model

Pre-model methods are self-contained and do not require a specific model architecture to be used. These methods include principal component analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE). In-model methods are interpretability techniques that are incorporated into the approach itself. Some methodologies are applied after creating a model and thus are referred to as post-model methods. These systems can build valuable observations about what a model gained throughout mentoring.

5.8 XAI's Evaluation Methods

Another significant aspect of the design process for XAI systems is the evaluation methodologies. Diverse metrics are prescribed to test the validity of the explanation for the desired purpose since explanations are created to meet a variety of interpretability aims.

5.8.1 Mental Model

A mental model is a depiction of how one perceives a system. according to cognitive psychology theories. Human–Computer Interaction (HCI) researchers examine users' conceptual models to see how much they know about intelligent systems in a variety of uses. One study looked at how individuals comprehend a smart grid system (Costanza et al. 2014), while another looked at how individuals comprehend and react to unpredictability in machine learning bus advent time predictions (Kay et al. 2016). Inquiring directly about the intelligent system's decision determining procedure is a good technique to examine users' understanding of intelligent systems. "Interviews, think-aloud, and self-explanations" provide useful data about an individual's thinking patterns and mental models when analyzed (Kim and Seo 1709).

5.8.2 Explanation Usefulness and Satisfaction

When evaluating explanations in expert systems, end-user convenience and the effectiveness of equipment justifications are also major determinants (Bilgic and Mooney 2005). To assess explanatory value for users, researchers use a variety of quantitative and qualitative criteria for comprehension, applicability, and adequacy of information (Miller 2019). Even though there are implicit methods of evaluating user acceptance (Hoffman 2018a, b), the evaluative assessment of contentment in descriptions, such as surveys and meetings, is a huge chunk of the literature.

5.8.3 User Trust and Reliance

User trust in expert machines is a cognitive and emotional factor that impacts whether users think a system is good or bad (Hoffman et al. 2013; Madsen and Gregor 2000). Swift trust (Meyerson et al. 1996), default trust (Merritt et al. 2013), and suspicious trust (Bobko et al. 2014) have all been used to describe the early individual faith and the trust-building with time.

5.8.4 Human-AI Task Performance

One of XAI's main goals is to assist end-users in becoming more prosperous in jobs using machine learning algorithms (Höök 2000). As a result, human-AI task productivity is a metric that applies to all user kinds. Users can adapt the intelligent system that caters to their requirements with the help of explanations. By delivering model interpretations, visual analytics tools also assist domain specialists in performing their responsibilities more effectively. Domain specialists can detect models and modify hyper-parameters to their particular data by evaluating model design, features, and machine output ambiguity. The necessity for model interpretation in a text (Hu et al. 2014; Liu et al. 2015; Wise et al. 1995) and multimedia (Bryan and Mysore 2013; Choo et al. 2010) assessment tasks has been investigated in visual analytics research.

5.8.5 Computational Measures

In the discipline of machine learning, computational measurements are commonly used to assess the correctness and completeness of classification strategies in terms of elucidating whatever the model has discovered. Instead of human subject investigations, computer tools should be used to determine the sincerity of explanations to the black-box model. The integrity of an improvised strategy in creating genuine justifications for model predictions is referred to as ad-hoc explainer integrity. As a result, a set of computational criteria for assessing the validity of produced interpretations, uniformity of interpretive outcomes, and authenticity of ad-hoc classification procedures to the initial black-box concept (Robnik-Šikonja and Bohanec 2018) have been developed.

5.9 XAI's Explanation Methods

In this section, we have listed some of the open-source methods used for the explanation.

5.9.1 Lime

An algorithm for faithfully explaining the outcomes of any encoder by estimating them effectively with an explainable fashion (Ribeiro et al. 2016).

5.9.2 Sp-Lime

An optimization strategy that identifies a group of relevant samples with justifications to handle the "trusting the model" problem (Ribeiro et al. 2016).

5.9.3 DeepLIFT

DeepLIFT is a deep learning recurrent estimation interpretive approach (Shrikumar et al. 1605). The ES attributes for a linear variant of the deep network are called DeepLIFT parameters. The use of DeepLIFT as an efficient framework for sampling-free estimation of ES attributes is motivated by this link. ES values can also be utilised to validate certain linearization choices made by DeepLIFT (Lundberg and Lee 1611).

5.9.4 Layer-Wise Relevance Propagation

Another method to approach the estimations of compositional networks (Bach et al. 2015). It is similarly an estimate of ES values, with the key contrast with DeepLIFT being the standard input used to estimate the impact of absent information (Lundberg and Lee 1611).

Some of the open-source methods used for the explanation are discussed above. Several typologies for classifying techniques for understandability have been suggested. Methods are classified as follows: characteristic value evaluation, reasoning from examples, and latent space traversal (Hase and Bansal 2020).

5.9.5 Characteristic Value Evaluation

Evaluation of characteristics value focuses on providing details on why the model uses particular aspects. The gradient-focused methods initially presented for vision by Simonyan et al. (1312), which could be transformed to be used with text data (Li et al. 1506), are the most prevalent of all these techniques. A variety of alternative methods for evaluating characteristic values all over data realms were suggested. LIME and Anchor are domain-agnostic approaches. Simple models, such as scattered linear models and instruction lists, are used in these techniques to estimate compound model actions relatively circling inputs. They demonstrate the evaluated results of explicitly interpretable characteristics on model results. What is "local" to input is described domain-specifically using an agitation allocation circling on that input?

5.9.6 Reasoning from Examples

Prototype models classify occurrences based on similarities to previously classified instances. Two papers on computer vision prototype models presented neural models that gain knowledge from prototypes correlating to image parts (Hase et al. 2019). These samples are being used to construct classifier features that are interpretable firsthand.

5.9.7 Latent Space Traversal

These methodologies navigate a model's dormant area to demonstrate how the model behaves when its input varies. Reaching the decision threshold in a classification setting could expose conditions required for a model's estimation of the native input. There are various techniques for developing a vision prototype (Joshi et al. 1806).

5.10 Explainable AI Stakeholders

A total of four stakeholders were considered in Explainable AI namely: Developers, Theorists, Ethicists, and Users (Preece et al. 1810).

5.10.1 Developers

Persons that are interested in developing AI applications. Many participants of this class provide their duties in industry, huge organisations, small and medium enterprises, and government, however, some are academicians or scholars who develop systems that can be used for a multitude of purposes, such as to help them with their work. Both the terms 'explainability' and 'interpretability' are used in this domain. Their key motivation for achieving explainability/interpretability is to improve the efficacy of their programs by offering aid to the testing process, troubleshooting, and review (Preece et al. 1810).

5.10.2 Theorists

Persons are enthusiastic about gaining knowledge and improving AI theory, notably in the domain of deep neural networks. Individuals are usually employed at scientific or corporate research facilities. Most are engaged practitioners; however, theorists

vary from developers in that their prime task is to improve the recent advancements in AI instead of providing real-world applications. The term 'interpretability' is more frequently used among theorists than 'explainability' (Preece et al. 1810).

5.10.3 Ethicists

Decision-makers, reporters, and reviewers who are worried about AI systems' justice, responsibility, and clarity. This group includes numerous computer engineers and technologists, but it also includes public scientists, attorneys, media people, bankers, and legislators. Members of this group also use the terms 'intelligibility' and 'legibility' in addition to 'explainability' and 'interpretability' (Preece et al. 1810).

5.10.4 Users

Individuals who employ artificial intelligence (AI) systems. The extensive mass of persons who participate in the growing compositions on AI explainability/interpretability belong to the first three circles, but this fourth one does not. Users require justifications to allow them to understand whether or not to respond in reaction to the outputs of the system, and/or to defend their decisions. This circle encompasses both real-life applications by users and engaged in procedures that are influenced by AI (Preece et al. 1810).

5.11 Applications

In this section, we have discussed some of the real-world implementations of Explainable Artificial Intelligence.

5.11.1 XAI for Training and Tutoring

After-action review (AAR) tools are common in combat training assistance, allowing trainees to reflect on their exercises and, presumably, learn from them. Mission statistics, a list of completed and failed objectives, and, in certain cases, a mission replay feature are all common aspects of these tools. It has been challenging for users of such AAR tools to grasp how their directions transfer into the behaviours of computer-controlled things with such constrained AAR tools due to increasingly complicated artificial intelligence (AI) in such training aids. Student users additionally have the drawback of having fragmented and sometimes incorrect domain knowledge: they

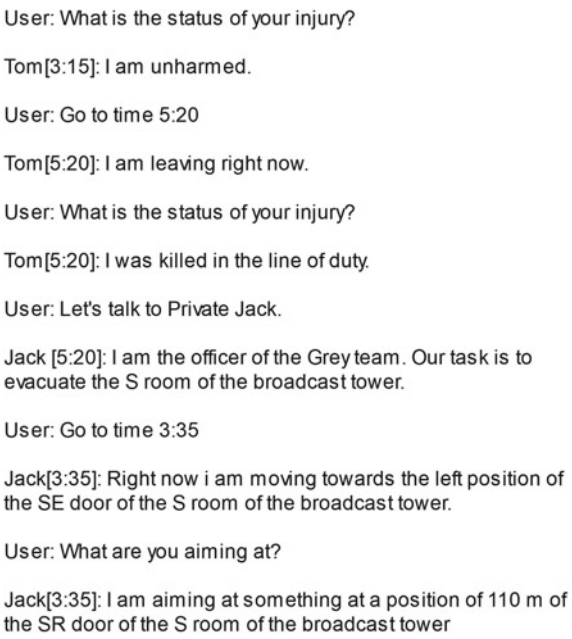

Fig. 5.3 XAI dialogue demo (Lane et al. 2005)

must not only learn new strategic knowledge and abilities but also comprehend emergent behaviors and their causes. They present a mechanism that allows entities in a strategic game to respond to questions regarding the actions. They explain continuing efforts to incorporate a smart instructor into the XAI framework and show how XAI may be used to deliver more meaningful after-action reviews (Lane et al. 2005). Figure 5.3 shows the XAI dialogue demo.

5.11.2 XAI for 6G

Explainable Artificial Intelligence (XAI) for 6G was described in detail, including public and legal incentives, criteria of understandability, efficiency versus understanding of research commutation, strategies to ameliorate explainability, and a framework for incorporating XAI into future wireless systems (Guo 2020). Figure 5.4 Illustrate XAI connectivity with UAV surveillance.

5.11.3 XAI for Network Intrusion Detection

Deep neural networks were utilized to detect network intrusions, and an explainable AI framework was suggested to give transparency to the machine learning pipeline

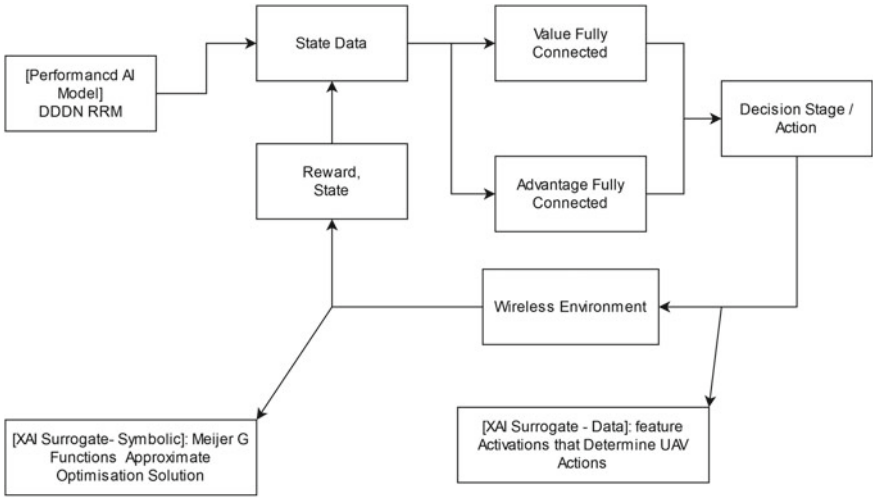

Fig. 5.4 Illustration of XAI connectivity with UAV surveillance (Guo 2020)

at every level. This is accomplished through the use of Explainable AI algorithms, which aim to make machine learning models not much like a black box by offering clarifications for why a prediction is made. Interpretations provide us with quantifiable data on which attribute impacts the likelihood of a cyber-attack or to what scale (Mane and Rao 2103).

5.11.4 XAI Planning as a Service

Explainable Planning can be implemented as a service, i.e., as a shell over a prevailing regulatory structure that allows the use of the existing planner to help answer incompatible questions. A prototype framework was introduced to help with this, as well as some instances of how a planner might be used to solve different types of incompatible issues. The key benefits and drawbacks of such an approach were then reviewed, followed by a questionnaire for Explainable AI Planning as a service that identified various possible research directions (Cashmore et al. 1908). Figure 5.5 illustrate the architecture of Explainable Planning as a service.

5.11.5 XAI for Prediction of Non-Communicable Diseases

Suggested a deep neural network framework based on Deep Shapley Additive Explanations (DeepSHAP) and supplied with a methodology for feature extraction for NCD prediction and explanation in the inhabitants. The DeepSHAP approach has

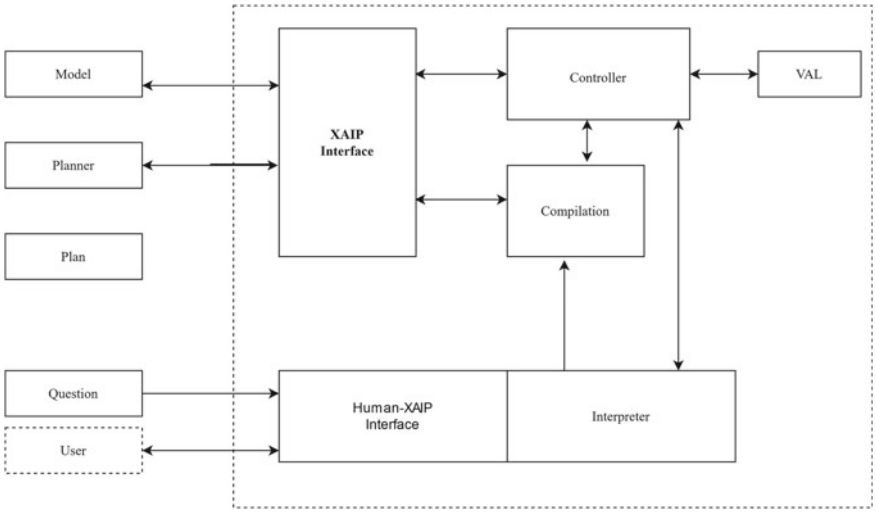

Fig. 5.5 Architecture for explainable planning as a service (Cashmore et al. 1908)

three constituents: first, the indicative characteristic is selected utilizing a flexible net-based immersed feature extraction approach; second, a deep neural network classifier is regulated with hyper-parameters and used to validate the algorithm with the chosen attributes batch; and third, the DeepSHAP approach provides two types of prototype elaboration. The developed scheme surpasses several existing models. Furthermore, the suggested model can aid the medical diagnosis of NCDs by offering a broad perception of changes in disease prospects at both the regional and international levels. The test findings show that key criteria that should have functioned to develop a confidence AI framework to differentiate between sufferers with COVID-19 signs and other sufferers may be interpreted using LIME (Davagdorj et al. 2021). Figure 5.6 shows the detection and understanding of noncommunicable illnesses using exploratory.

5.11.6 XAI for Scanning Patients for COVID-19 Signs

The objective of this project is to create a deep learning-based concept that can accurately recognize COVID-19 sufferers on a CT scan and a chest X-ray image collection. Eight alternative deep learning algorithms were updated and evaluated on two datasets: one with four hundred CT scan images and the other with four hundred chest X-ray images in this study. With a Ninety Five percent confidence interval, NasNetMobile surpassed those other models entitle of performance on CT scan (81.5–95.2%) and chest X-ray (95.4–100%) image datasets. In addition, the model's interpretability is simplified using Local Interpretable Model-agnostic Explanations

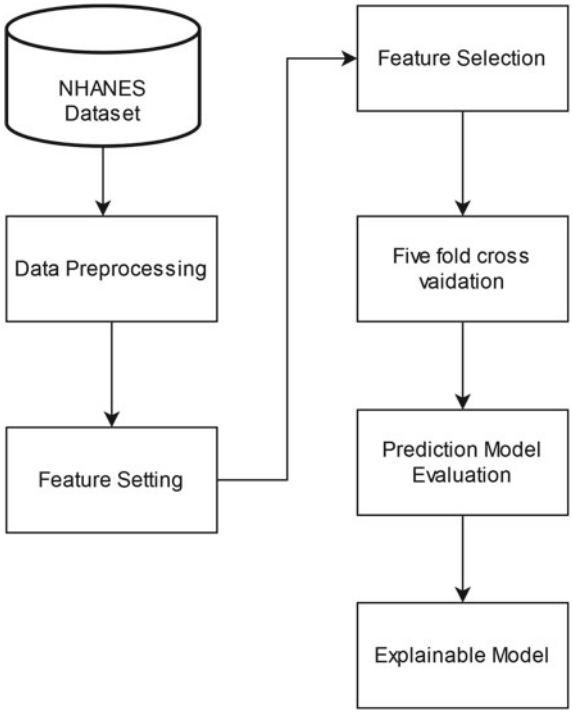

Fig. 5.6 Detection and understanding of noncommunicable illnesses using exploratory (Joshi et al. 1806)

(LIME) (Ahsan et al. 2007). Figure 5.7 displays the complete process flow of the study.

Many frameworks established in the subject of Explainable Artificial Intelligence are discussed in this section.

A design and assessment framework is suggested for end-to-end explainable artificial intelligence systems composition, as demonstrated through a prototype and suggestions, that integrates design goals with evaluation techniques (Mohseni et al.

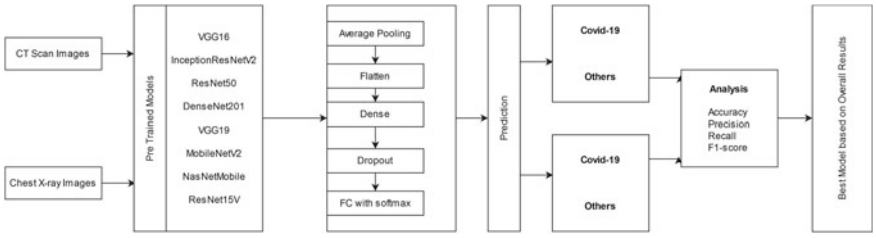

Fig. 5.7 Complete study process flow (Preece et al. 1810)

2021). A new functional explanation framework was developed, which differed from the majority of XAI's previous explanation frameworks. A functional explanation is evaluated using three criteria: correlation, completeness, and complexity. The correlation ensures that the explainer's information can bridge the gap between the explainer and the explainee. The level of correlation is measured using Pearl and Mackenzie's ladder of causation. Explainer information should be complete to describe the underlying systems. Six requirements were provided for a detailed description.

The explanation should be as straightforward as possible for the explainee, albeit this criterion is less crucial than correlation and completeness. The current state of XAI approaches was then assessed using these three variables. According to the data, the rule-extraction approach could provide the greatest standard of explanation on the hypothetical level in terms of correlation (Cui et al. 2019).

An explainable AI framework for imaging air filters was described, which can be expanded to other imaging applications including damage, flaws, or abnormalities. A feedforward neural network was used to build a quantifiable length pseudometric, which is then used as an exclusionary predictor for spatial bootstrap unorganized picture data. After that, this classifier is utilized as a forecaster in a Bayesian inference/regression model to forecast air filter erode degree with a 95% confidence level. The AI model's explainability was attained by using a "twin" model having understandable insights components and correlating them with the AI model's components. This explainable AI model could be used in a range of applications that use artificially produced structured and non—structured imagery in predictive nurturing and health management (Krishnamurthy et al. 2020).

A normative paradigm for Explainable Artificial Intelligence was designed to solve the Black Box Problem. This normative paradigm not only demonstrates the utility of the analytic approaches being produced in Explainable AI but rather they possess some restrictions. Diagnostic classification and feature-detector-identification approach, like input heatmapping, function best when the system's variables can be evaluated logically. Although semantic applicability is typically desirable, it is not always necessary, according to the framework study (Zednik 2021).

A novel theoretical framework to unite exploration and methods evolved in the domain of explainable AI (XAI) was proposed to solve the increasing diversity and absence of cooperation on what defines a comprehensible or explainable model was introduced to address the growing diversity and deficiency of agreement on what establishes an explainable or interpretable model. Two key concepts, "explanation" and "interpretation," underpin this paradigm. These concepts are further framed within a generic workflow that restricts other crucial semantics such as input/output realms and creates a distinction between low-level mathematical concepts and the high-level, human-understandable arena of non-functional constraints. Furthermore, the framework was used to demonstrate how it might aid in the evaluation of existing XAI approaches, demonstrating the amount by which each addresses distinct components of the explainability process (Palacio et al. 2021).

5.12 Possible Research Ideology and Discussions

In this section, we look at new work opportunities for XAI and spot possible scientific pathways that can be adopted to tackle them efficiently shortly.

An examination of the XAI literature reveals that, given the relatively latest and multidisciplinary nature of XAI, an organized approach to core ideas has still not been properly developed. The latest research, on the other hand, is supplying more detailed concepts that are causing a significant influence and forming the domain.

The field's infancy is indeed noticeable in the strategies to develop and construct XAI frameworks and outcomes that have been motivated by devs. Using XAI to reduce prejudices, guarantee social responsibility, and equity requires much data preprocessing and modeling techniques as it does demonstrate that it is properly considered. As a developer resource, XAI can be extremely useful in ensuring that a model relies on causative factors rather than commonalities, that data is equitably spread, and that the functionalities used are appropriate. Possessing the subject matter expertise, that a practitioner may have had in his domain is never sensible, but it is supposed to create sound models. Again, more cross-disciplinary research in XAI is required to make sure that it can collude across competent realms and acquire competence from all relevant areas. When it comes to the use of XAI in Artificial Based architectures, the main discussions show that organizations must take precautions when selecting an XAI approach for addressing a particular requirement. To produce XAI concepts that fulfill regulators and enforce adherence, particularly in healthcare and finance, showed a strong tier of involvement, as this is a considerable challenge in incorporating ML in these immensely governed aspects. The writings strongly imply that XAI will reduce this restriction; even so, no considerable evidential research has been conducted to demonstrate how XAI may fulfill regulators' demand to maintain adherence or undertake evaluations/audits.

How to validate interpretations and associated paradigms is again an evolving trend in XAI. On the one side, investigators should guarantee that sentient explanation is accepted, while on the other side, frameworks must display details as it is, without obscuring the metrics and instead of constructing convincing outcomes.

Conceptual debates about sentient interpretations are presently influenced by sociological viewpoints, with little regard for what tech is proficient in providing. This strengthens the social-technical duality, resulting in a highly fragmented XAI domain centered on either how we as beings analyze interpretations or how one can technologically retrieve features from complicated models. While the present state is warranted by the domain's infancy, more research is necessary to clarify the various XAI stakeholders' necessities and how they can be met with much more aimed interpretations than the two presiding clusters of devs or users.

Whilst the research discussion is far from conclusive, it has identified two directions for future XAI studies:

- The need to contribute to the key parties and their various explanatory requirements. This research direction emphasizes the significance of examining the micropolitics of XAI in groups and its impact on tasks.

- The demand for a systematic perspective in researching XAI, keeping in mind both the sociotechnical elements of XAI, as well as the procedure and result factors of XAI, as well as the credible and narrative components of XAI. This research path underlines the value of conceptualizing and scientifically understanding the multidimensionality of XAI a novel type of socio-technical system, as well as its significance for AI techniques in society and business.

5.13 Conclusion

In the last decagon, as principled considerations, laws, and the need to govern these models have grown, the desire to unlock the renowned 'AI black box' has gained a lot of traction. On the XAI, we conducted a rigorous literature review. We begin by determining the evaluation procedures. Second, the techniques of explanation were discussed, followed by the stakeholders, applications, and frameworks. From the viewpoint of the stakeholders, it is considered that this field necessitates more investigation. It is also brought to our knowledge that black box approaches are unsuitable for some applications, such as in the health world, where a system's incorrect calls might be extremely dangerous. Explainability was also emphasized as a requirement for resolving legal issues that have arisen as a result of the expanding use of AI systems. We anticipate that it will enhance community awareness of the primary categories among scholars (methods, applications, frameworks, etc.).

References

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

Ahsan, M.M., Gupta, K.D., Islam, M.M., Sen, S., Rahman, M., Hossain, M.S., et al.: Study of different deep learning approach with explainable AI for screening patients with COVID-19 symptoms: using CT scan and chest X-ray image dataset (2020). arXiv preprint arXiv:2007.12525

Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: results from a systematic literature review. In: 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019 (2019)

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. fusion **58**, 82–115 (2020)

Attaran, M., Deb, P.: Machine learning: the new 'big thing' for competitive advantage. Int. J. Knowl. Eng. Data Min. **5**, 277–305 (2018)

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**, e0130140 (2015)

Bilgic, M., Mooney, R.J.: Explaining recommendations: satisfaction versus promotion. In: Beyond Personalization Workshop, IUI (2005)

Bobko, P., Barelka, A.J., Hirshfield, L.M.: The construct of state-level suspicion: a model and research agenda for automated and information technology (IT) contexts. Hum. Factors **56**, 489–508 (2014)
Bryan, N., Mysore, G.: An efficient posterior regularized latent variable model for interactive sound source separation. In: International Conference on Machine Learning (2013)
Cashmore, M., Collins, A., Krarup, B., Krivic, S., Magazzeni, D., Smith, D.: Towards explainable AI planning as a service (2019). arXiv preprint arXiv:1908.05059
Choo, J., Lee, H., Kihm, J., Park, H.: iVisClassifier: an interactive visual analytics system for classification based on supervised dimension reduction. In: 2010 IEEE Symposium on Visual Analytics Science and Technology (2010)
Costanza, E., Fischer, J.E., Colley, J.A., Rodden, T., Ramchurn, S.D., Jennings, N.R.: Doing the laundry with agents: a field trial of a future smart energy system in the home. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (2014)
Cui, X., Lee, J.M., Hsieh, J.: An integrative 3C evaluation framework for explainable artificial intelligence (2019)
Davagdorj, K., Bae, J.-W., Pham, V.-H., Theera-Umpon, N., Ryu, K.H.: Explainable artificial intelligence based framework for non-communicable diseases prediction. IEEE Access **9**, 123672–123688 (2021)
Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017). arXiv preprint arXiv:1702.08608
Gautam, A., Chatterjee, I.: An overview of big data applications in healthcare: opportunities and challenges. In: Knowledge Modelling and Big Data Analytics in Healthcare, pp. 21–36 (2021)
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) (2018)
Guo, W.: Explainable artificial intelligence for 6G: improving trust between human and machine. IEEE Commun. Mag. **58**, 39–45 (2020)
Hase, P., Bansal, M.: Evaluating explainable AI: which algorithmic explanations help users predict model behavior? (2020). arXiv preprint arXiv:2005.01831
Hase, P., Chen, C., Li, O. Rudin, C.: Interpretable image recognition with hierarchical prototypes. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2019)
Hoffman, R.R., Johnson, M., Bradshaw, J.M., Underbrink, A.: Trust in automation. IEEE Intell. Syst. **28**, 84–88 (2013)
Hoffman, R.R.: Theory → concepts → measures but policies → metrics. In: Macrocognition Metrics and Scenarios, pp. 3–10. CRC Press (2018a)
Hoffman, R.R.: Theory concepts measures but policies metrics. In: Macrocognition Metrics and Scenarios, pp. 3–10. CRC Press (2018b)
Höök, K.: Steps to take before intelligent user interfaces become real. Interact. Comput. **12**, 409–426 (2000)
Hu, Y., Boyd-Graber, J., Satinoff, B., Smith, A.: Interactive topic modeling. Mach. Learn. **95**, 423–469 (2014)
Joshi, S., Koyejo, O., Kim, B., Ghosh, J.: xGEMs: generating examplars to explain black-box models (2018). arXiv preprint arXiv:1806.08867
Kay, M., Kola, T., Hullman, J.R., Munson, S.A.: When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (2016)
Kim, J., Seo, J.: Human understandable explanation extraction for black-box classification models based on matrix factorization (2017). arXiv preprint arXiv:1709.06201
Krishnamurthy, V., Nezafati, K., Stayton, E., Singh, V.: Explainable AI framework for imaging-based predictive maintenance for automotive applications and beyond. Data-Enabled Discov. Appl. **4**, 1–15 (2020)
Kumar, A., Chatterjee, I.: Data mining: an experimental approach with WEKA on UCI Dataset. Int. J. Comput. Appl. **138** (2016)

Kumar, D., Mehta, M.A., Chatterjee, I.: Empirical analysis of deep convolutional generative adversarial network for ultrasound image synthesis. Open Biomed. Eng. J. **15** (2021)
Lane, H.C., Core, M.G., Van Lent, M., Solomon, S., Gomboc, D.: Explainable artificial intelligence for training and tutoring (2005)
Li, J., Chen, X., Hovy, E., Jurafsky, D.: Visualizing and understanding neural models in NLP (2015). arXiv preprint arXiv:1506.01066
Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue **16**, 31–57 (2018)
Liu, M., Liu, S., Zhu, X., Liao, Q., Wei, F., Pan, S.: An uncertainty-aware approach for exploratory microblog retrieval. IEEE Trans. Vis. Comput. Graph. **22**, 250–259 (2015)
Lundberg, S., Lee, S.-I.: An unexpected unity among methods for interpreting model predictions (2016). arXiv preprint arXiv:1611.07478
Madsen, M. Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems (2000)
Mane, S., Rao, D.: Explaining network intrusion detection system using explainable AI framework (2021). arXiv preprint arXiv:2103.07110
Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I trust it, but I don't know why: effects of implicit attitudes toward automation on trust in an automated system. Hum. Factors **55**, 520–534 (2013)
Meyerson, D., Weick, K.E., Kramer, R.M., et al.: Swift trust and temporary groups. In Trust in Organizations: Frontiers of Theory and Research, vol. 166, p. 195 (1996)
Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
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. (TIIS) **11**, 1–45 (2021)
Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proc. Natl. Acad. Sci. **116**, 22071–22080 (2019)
Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., Dengel, A.: Xai handbook: towards a unified framework for explainable AI. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
Preece, A., Harborne, D., Braines, D., Tomsett, R., Chakraborty, S.: Stakeholders in explainable AI (2018). arXiv preprint arXiv:1810.00184
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 (2016)
Robnik-Šikonja, M., Bohanec, M.: Perturbation-based explanations of prediction models. In: Human and Machine Learning, pp. 159–175. Springer (2018)
Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**, 206–215 (2019)
Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: learning important features through propagating activation differences (2016). arXiv preprint arXiv:1605.01713
Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps (2013). arXiv preprint arXiv:1312.6034
Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)
Villata, S., Boella, G., Gabbay, D.M., Van Der Torre, L.: A socio-cognitive model of trust using argumentation theory. Int. J. Approx. Reason. **54**, 541–559 (2013)
Wise, J.A., Thomas, J.J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. Crow, V.: Visualizing the non-visual: spatial analysis and interaction with information from text documents. In: Proceedings of Visualization 1995 Conference (1995)
Zednik, C.: Solving the black box problem: a normative framework for explainable artificial intelligence. Philos. Technol. **34**, 265–288 (2021)