On the Different Concepts and Taxonomies of eXplainable Artificial Intelligence

Arwa Kochkach[1(✉)], Saoussen Belhadj Kacem[2], Sabeur Elkosantini[2], Seongkwan M. Lee[3], and Wonho Suh[4]

[1] Higher Institute of Management of Tunis, University of Tunis, Tunis, Tunisia
arwa.kochkache@gmail.com
[2] Faculty of Economics and Management of Nabeul, University of Carthage, Tunis, Tunisia
saoussen.belhadjkacem@fsegn.u-carthage.tn, Sabeur.elkosantini@fsegn.ucar.tn
[3] Colleage of Engineering, United Arab Emirates University, Al Ain, United Arab Emirates
MarkLee@uaeu.ac.ae
[4] Hanyang University, ERICA Campus, Seoul, South Korea
wonhosuh@hanyang.ac.kr

Abstract. Presently, Artificial Intelligence (AI) has seen a significant shift in focus towards the design and development of interpretable or explainable intelligent systems. This shift was boosted by the fact that AI and especially the Machine Learning (ML) field models are, currently, more complex to understand due to the large amount of the treated data. However, the interchangeable misuse of XAI concepts mainly "interpretability" and "explainability" was a hindrance to the establishment of common grounds for them. Hence, given the importance of this domain, we present an overview on XAI, in this paper, in which we focus on clarifying its misused concepts. We also present the interpretability levels, some taxonomies of the literature on XAI techniques as well as some recent XAI applications.

Keywords: EXplainable Artificial Intelligence · Interpretability · Explainability · Post-hoc explanation techniques

1 Introduction

Recently, the sophistication and advancement of the Artificial Intelligence (AI)-powered systems has increased exponentially. Indeed, it reached a scope that "almost no human intervention is required for their design and deployment" [2]. However, although these models exhibit high performance, many of them are opaque in terms of explainability, i.e. they are not able to provide an explanation of their outputs. In this context, many Machine Learning (ML) models are considered as black boxes such as Artificial Neural Networks (ANN) or gradient boosting machines [22].

A. Bennour et al. (Eds.): ISPR 2023, CCIS 1941, pp. 75–85, 2024.
https://doi.org/10.1007/978-3-031-46338-9_6

It is significant to note that there exist numerous cases and situations where the explanation of the AI application is not necessary or needed at all. This is especially true when the problem under study and treatment is well represented, well-known, complete and its consequences are not critical (e.g. movie recommendation or mail sorting) [8]. However, explanations are crucial for the user to comprehend and trust the decisions yielded from a system for many other critical cases [14]. This is mainly true when these latter have an effect on humans' lives (e.g. healthcare, medicine, defense or law). Moreover, as declared by Zhu et al. [30], "humans are reticent to adopt techniques that are not directly interpretable, tractable and trustworthy". Hence, most of the community members stand, nowadays, in front of the barrier of "explainability". Paradigms underlying this issue fall within the so-called eXplainable Artificial Intelligence (XAI) field. Gunning [13] states that XAI "will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners". Nevertheless, although it may not necessarily occur in many datasets [27], there is a widespread belief that there exists a trade-off between the interpretability of a model and its performance (e.g. predictive accuracy) [9]. For this reason, XAI's goal is the creation of more interpretable ML systems while preserving their high level of learning performance in order to avert the limitation and the sacrifice of the effectiveness of the current AI-powered systems [2].

Within the field of XAI, there is a lot of concepts which need to be well identifed and understood in order to make correct and flawless insights. Indeed, many of them are interchangeably misused in the literature. The most notable ones are "interpretability" and "explainability". This issue, mainly, obstructed the foundation of common grounds for XAI's nomenclatures. For this reason, it is crucial to make a clear distinction between them. Therefore, this will be the goal of this paper. In fact, our aim is to present a basic overview and to summarize the main nomenclature frequently utilized in XAI community. We also clarify the distinctions and similarities among them.

The remaining of this article is structured according to the following steps. Section two is dedicated for the definition of some XAI-related concepts as well as giving a clear line to distinguish between them according to our proposed criteria. Then, the levels of interpretability of a model are presented in Sect. 3. Section 4 will contain our proposed taxonomy to categorize XAI techniques within the ML field based on different ones from the litterature. Lastly, some of the application domains of XAI will be shown in Sect. 5. And then, we end this paper with a conclusion.

2 On the Concepts of eXplainable Artificial Intelligence

The term "explainability", being relevant in the literature, gave rise to the direction of XAI [29]. However, there is some ambiguity in this field with regard to terminology [6]. This was well outlined by Lipton [20] who notes that also "the term

interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way".

Indeed, in the literature, there were several attempts to define the concepts of "explainability" and "interpretability". Despite that, generally speaking, there is no consensus within the ML community on their definitions [8,20,22]. Hence, many researchers consider them as the same [6,7] and they usually use them interchangeably in their works "in the broad general sense of understandability in human terms" [12]. According to our researches, the most popular definition, considered for both of them, adopted for numerous authors [2,6,9,12,19,22] is that of Doshi-Velez and Kim [8] stating that they represent "the ability to explain or to present in understandable terms to a human" . Besides, we note that there exist other terms, in the literature, such as "intelligibility" [5,21], "understandability" [20] and comprehensibility [9,11] that are considered synonymous to interpretability and they are often used interchangeably also [22].

However, many other authors argue that there are important reasons to distinguish between the concepts of "explainability" and "interpretability" [12,22,27]. Hence, they try to identify their differences in their works. Indeed, there is no concrete mathematical definition for both of the concepts i.e. they have not been measured by metrics [19]. Nevertheless, many tries have been made in order to make them clear. For example, authors in [12] assume that "explainable models are interpretable by default but the reverse is not always true". Moreover, in [4], authors discern that "interpretability is usually used in terms of comprehending how the prediction model works as a whole" while explainability "is often used when explanations are given by prediction models that are incomprehensible themselves". We believe that the most relevant distinction is made by Rudin [27] who marks a clear line between explainable and interpretable techniques. He states that the latter ones "focuse on designing models that are inherently interpretable"; whilst the former ones "try to provide post hoc explanations for existing black box models". For the sake of convenience, in our paper, we adhere this dissimilarity. We believe that this is the most pertinent one with regards to the field of XAI. Hence, we present these concepts more thoroughly in this section according to this belief.

2.1 Interpretability

Interpretability is defined by Miller [23] as "the degree to which a human can understand the cause of a decision". The property of the interpretability of a ML model stems from the model itself, i.e. it is intrinsic and built-in. In other words, this concept regards a model that is clear by design and does not need to be explained by another technique. Kim et al. [17] describe interpretability, in a more accurate way, as "the degree to which a human can consistently predict the model's result". Based on the above, interpretability is then chiefly attached to the intuition behind a model's outcome [1]. Thus, it is higher if it is easier for a human user to identify the causes and the effects the inputs have on the outputs of the model, i.e it is easier to trace back why a given prediction was output [6,19].

We note that, although being intuitive, the aforementioned definitions clearly lack mathematical rigour and formality i.e. there is no mathematical definition of interpretability [1,6,20]. In fact, none of them is specific or restrictive enough to enable formalization. Besides, many authors assume that "interpretability is a very subjective concept" hence it is not that easy to state it formally [28].

An interpretable model is a ML model that is interpretable by design. It has inherent/ in-model interpretability in it i.e. the interpretability process takes place during the building of the ML model. It is also called an "interpretable by nature" [4], intrinsically interpretable [6], transparent [2], or white-box model [19]. This can be achieved if the model is simple enough by nature or by imposing some relevant constraints on the complexity of the ML model to be developed. These constraints can be causality, monotonicity, sparsity, additivity, and/or other desirable properties [6]. The model can be structured to reflect some physical constraints coming from the domain knowledge too [27]. This can also be done via the hybridization or the mapping of the black-box system with a more interpretable ML model (a white-box twin) [2]. For example, in [16], a Deep Neural Network and a Case Based Reasnoning model (a k-Nearest Neighbors) are paired with a view to enhance the interpretability of the former while maintaining its high level of accuracy. Neural networks can also be explained by mixing them with fuzzy systems giving rise to the famous interpretable Neuro-Fuzzy Systems [26]. There exists also the process of the "deep formulation" of the classical ML models [2]. A prominent example of the latter model improved by its Deep Learning (DL) counterpart is Deep KNN (DkNN) [25].

2.2 Explainability

The issue of explainability was not present in the last miscellany of AI techniques (namely rule based models and expert systems) and it came to light recently. This problem arises and explainability becomes required when there is some degree of incompleteness in a problem formalization making direct optimisation and validation impossible [4,8,22] in some concrete situations such as Scientific Understanding, Safety and Ethics [22]. Incompleteness denotes that "there is something about the problem that cannot be sufficiently encoded into the model". It is different from uncertainty which hints at "something that can be formalized and handled by mathematical models" [8]. In fact, for some incomplete problems or prediction tasks, the resulting output representing the prediction solely (the "what") it is not enough. Hence, the model should also explain and makes clear how it came to this prediction (the "why") [6].

Explainability establishes an interface of interaction between humans and a decision-making model (the ML model). Hence, it has to be an accurate proxy of this latter while being well comprehensible to humans. However, each users group may vary in their background knowledge and may have a preferred explanation type that is able to communicate information in the most effective way [14]. For this reason, the concept of "audience" is being the cornerstone of XAI [2]. Hence, authors in [2] argue that its definition must be rephrased to reflect this dependence explicitly as follows: "Given a certain audience, explainability refers

to the details and reasons a model gives to make its functioning clear or easy to understand" [2]. Put it differently, explainability is "the ability a model has to make its functioning clearer to an audience" [2]. Indeed, authors in [19] point out that it concerns the internal logic and procedures which are executed while training the model and making decisions. Hence, its goal is to describe them in a way that is comprehensible to humans [12]. A more relevant definition is given by Rudin [27] which assumes that "explainable ML focuses on providing explanations for existing black box (opaque) models by means of training another surrogate model post hoc". These post hoc explainability techniques (also called "explanation methods" or "post-modeling explainability") are specially devised to explain/ improve the interpretability of an existing opaque ML model, which do not meet any criterion allowing to deem it interpretable, after its building and training by analyzing it [6,24]. For this reason, they have to be able to summarize the reasons for its behavior, produce insights and communicate understandable information about the causes of their decisions (predictions) and gain the trust of users [12]. Many authors argue that this notion frequently depends on the domain of application [11,27,28], the user [6] (his abilities, expertise, preferences and expectations [9,14]), the context that depends on the task and other contextual variables [14]. Therefore, an all-purpose definition might be unnecessary [22] or infeasible [27].

2.3 Explainable vs Interpretable Machine Learning

As can be seen in the previous subsection, there are notable differences among XAI concepts, mainly Interpretability and Explainability. In table 1, we propose some criteria that help to distinguish between them with the intent to draw a clear line discerning between the two of them.
W ithin the XAI domain, there exist several other concepts which often appear in the literature such as:

- Transparency, which can be considered as the obverse of "black-box-ness", is "a feature that a model can feature by itself" [2]. It can also be defined as "the search for a direct understanding of the mechanism by which a model works" [20]. Authors in [9] state that a transparent model is both interpretable and explainable. Others use it interchangeably with interpretable models [2]. In this review we adopt that transparency refers to the highest level of interpretability i.e. simulatability.
- Understandability which is also ofen used interchangeably with interpretability and intelligibility [5,20–22]. Authors in [2] state that understandability is a two-sided matter: model understandability vs. human understandability. Model understandability is" the characteristic of a model to make a human user understand its functioning without any need for explaining its internal structure". However, "human understandability measures the degree to which a human user can understand a decision made by a ML model" i.e. relies on the capability of the specific audience.

2.4 Explanation

As we noted previously, the concept of explanation is relevant in the field of XAI since explainability is clearly associated with it. Miller [23] presents a simple goal-oriented definition of an explanation stating that "it is the answer to a "Why" question". Within the XAI community, an explanation is defined as "the means by which a ML model's decisions are explained" [6] in a humanly understandable fashion. Taking the abovementioned definitions into consideration, an explanation is deemed to be pertinent if it rejoin to the needs and goals of the user[4] (providing as much information as possible while being as short as possible), allows a tradeoff between explainability and completeness (descriptions with a high level of details) [12] and convincing to the user (he accepts it) [4].

There exists a wide range of post-hoc explanation techniques in the literature. Among them we can mention Feature Relevance Explanation techniques which measure the influence or importance which every managed input feature (variable) has on the output coming out of the opaque ML model. Another example can be the Local Explanations which process first by segmenting the solution space then generating explanations for some less complex relevant subspaces. Explanation by Simplification functions by rebuilding a whole simple (less complex) new system to explain the original one while preserving a similar performance score [2]. Then we can cite Visual Explanation techniques which

Table 1. Interpretable vs. Explainable Machine Learning.

Criterion	Interpretable	Explainable
Characteristic of the model	Passive characteristic: "the level at which a given model makes sense for a human observer" [2].	Active characteristic: an external action taken by a model in order to clarify its internal functioning [2].
Means of explanation	Models are interpretable by design [2].	Models are explained via external XAI techniques [2].
Level of the explanation	Intrinsic/ Inherently interpretable: explanations exist in the model itself [2,22,27].	Post hoc: the explanation is given by another model "after" the training of the existing black-box one [27].
Which question it raises?	"How does the model work?" [20]	"What else can the model tell me?" [20]
Extent of the explanation	Is not necessarily a model that human users are able to comprehend its internal logic and processes [19].	Deal with the internal logic and procedures of a black-box system [19].
Focus	Model-centric [10]	Subject-centric [10]
Mode of explanation	Users can mathematically analyze the mappings [7]	Models should output symbols, rules or figures with their prediction to help the user to understand the rational behind the input-output mappings being made [7].
Need for the explanation	Used to comprehend how the model works superficially and as a whole [4].	Used when a model is incomprehen--sible itself and its formalization is incomplete [4].

goal is the visualizing the ML model's comportment, Explanations by Example which "extractes representative examples that grasp the inner relationships and correlations found by the model" and Text Explanations [3].

3 Levels of Interpretability

An AI model can feature different levels of Interpretability [2]. Hence, within Interpretable models, three levels are contemplated with regards to the scope of interpretability: the portion of the prediction process targeted for explanation [6]. This gives rise to three different categories of AI models: simulatable, decomposable and algorithmically transparent models [20]. Authors in [2] introduce thoroughly these three classes as follows:

1. *Algorithmic transparency* is the lowest level of interpretability. "It deals with the ability of the user to understand the process followed by the model to produce any given output from its input data" [2]. It also refers to the way works the algorithm which generates or learns the ML model from the data [24]. Thus, it only requires acquaintance with the model's algorithm and it does not concern the managed data. Thus, for a model to be algorithmically transparent, it has to be "fully explorable by means of mathematical methods and analysis" [2].
2. *Decomposability* is the second level of interpretability (on a modular level). It refers to the ability to give an explanation about each part of a ML model such as inputs or parameters. Hence, it responds to the question "how do parts of the model affect predictions" [24]. However, this property does not feature for every ML model since it requires every input used to train it to be readily interpretable which does not apply for cumbersome features for example.
3. *Simulatability* (Global Model Interpretability [6], transparency) is the highest level of interpretability. It denotes the capability a model have to be wholly and strictly simulated by they user. Moreover, Lipton [20] defines it as "the ability to comprehend the entire model at once". To do so, the trained model and the knowledge of the algorithm (each of the learned components such as parameters, weights, etc) as well as a holistic view of the data features are needed [6]. Therefore, it is very difficult to attain this level in practice [24]. Honegger [15] acknowledges that, for a model to fulfill this criteria, it must be simple enough.

4 Taxonomies of Interpretable and Explainable Machine Learning

ML interpretability and explainability techniques can be classified according to different criteria. In the following, we review some of them.

4.1 Model-Specific vs. Model-Agnostic Techniques for Post-hoc Explainability

This is an important criterion distinguishing between two different algorithmic approaches as follows.

1. *Model-agnostic methods* are techniques which are designed to be linked seamlessly to any ML model (black box or not) with the intention of extracting knowledge about the decision-making procedure [2]. Hence, they count just on analyzing pairs of input and output [6,9] and they cannot have access to the internal representations or the inner working process of the black-box ML model such as weights or structural information [2,24].
2. *Model-specific explanation methods* are tailored or specifically designed for a specific ML model or one class of them [22] since they are based on some particular model's internals [24] and it uses idiosyncrasies of its representation [9]. Hence, they cannot be directly plugged to any other model [2].

4.2 Post-hoc Explainability Techniques for Shallow vs. Deep Machine Learning

Authors in [2] propose a taxonomy of the post-hoc explainability techniques which divide the literature into two main categories: techniques devised for shallow ML models and others devised for DL models. In the following, both of them are presented.

1. *Post-hoc explainability in shallow ML models:* Shallow models collectively refer to all ML models which structure is not layered following neural processing units i.e. they have a relatively straightforward structure. Within these models, there are the strictly interpretable (transparent) ones (e.g. Decision Tree or K-NN) and others that have more sophisticated or complex learning algorithms. Hence, the latter ones require additional post-hoc explanation (e.g. Support Vector Machines or Tree Ensembles)[2].
2. *Post-hoc explainability in deep learning:* The most common DL models are multi-layer neural networks (MLNN), Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). They are efficient and widely used thanks to their great capabality of deducing such complex connections among different and numerous variables. However, their structures are extremely complex and difficult to understand which makes them always considered as black-box models. Hence, they require explainability techniques to make their functioning clear for the user [2].

Figure 1 provides one possible proposed taxonomy of the explainable and interpetable models that we exposed in this section. It is noteworthy that this taxonomy is not a disjoint partition of the suite of all the techniques. In fact, almost all the discussed criteria in this sections are related to each other in some way. In fact, the interpretability of interpretable models lies in the core of the model (intrinsic) and is present while training it (In-model) hence it is always model

specific. This interpretability can be global touching the whole model or local in some of its parts. In deep models, the structure is always complex and hard to understand. Hence, they always need post-hoc techniques to explain them which are resorted to after the training of the model (Post-model). These latter techniques can be either model agnostic or model specific. We note that, although there are few model-specific techniques that are deployed post-hoc, most of them are achieved through intrinsically interpretable models. In an analogous way, most of post-hoc methods are separated from the models and, hence, they are model-agnostic [6].

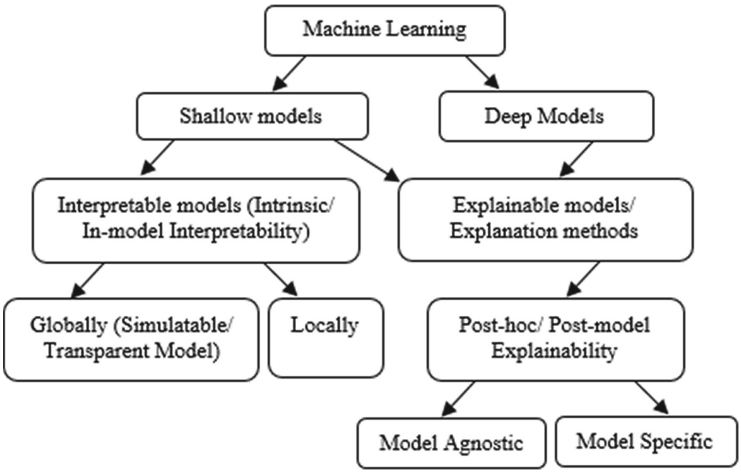

Fig. 1. An overall taxonomy of Interpretable and Explainable ML techniques.

5 Application Domains of eXplainable Artificial Intelligence

At present, XAI is becoming paramount and needed in many application domains. Mainly, when the result (prediction, assessment, decision or classification) of the ML model affects human's life, its explanation to the user becomes crucial to enhance the faithfulness of this model. Indeed, it is always needed in medicine and healthcare. For example, it was applied for the predicting of pneumonia risk [5] and surgical effort in advanced stage epithelial ovarian cancer [18]. Moreover, XAI was also employed in military simulations and computer games. For example, in [29], it was used to provide explanations utilized by a training framework specially designed for the U.S. Army for small-unit tactical behavior. Besides, XAI can be used in some non-critical domains like entertainment. For example, in [30], it was used for games designers to help them better employ AI and ML in their design tasks through co-creation.

6 Conclusion

Our paper has overviewed the field of XAI which is being more and more identified as an essential requirement, in some real-life applications, within the community of AI. Our research study has elaborated on this theme by, first, shedding light on some XAI-related concepts and by clarifying their interchangeable misuse. This is mainly done for interpretable models which are interpretable by design and explainable models which need an external XAI technique to explain them after their training. Secondly, we presented the three levels of interpretability. Thereafter, a proposed overall taxonomy of recent literature dealing with XAI techniques was presented according to different criteria and then we ended by overviewing some application domains of the field.

Although being widely manipulated in the literature, XAI techniques are fewly applied in some critical domains such as law, defense and security. In fact, most of the authors are focusing on developing new XAI techniques rather then applying the existent ones in such domains. Hence, they need to be further used in order to make these systems more trustworthy and reliable for the user.

Acknowledgements. This research work was supported and funded by Data4Transport, a Tunisian-South Korean research project. The authors would like to thank all personnel involved in this work.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Arrieta, A.B., et al.: Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion **58**, 82–115 (2020)
3. Bennetot, A., Laurent, J.L., Chatila, R., Díaz-Rodríguez, N.: Towards explainable neural-symbolic visual reasoning. arXiv preprint arXiv:1909.09065 (2019)
4. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. **70**, 245–317 (2021)
5. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730 (2015)
6. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019)
7. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
9. Došilović, F.K., Brčić, M., Hlupić, N.: Explainable artificial intelligence: a survey. In: 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), pp. 0210–0215. IEEE (2018)
10. Edwards, L., Veale, M.: Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. Duke L. Tech. Rev. **16**, 18 (2017)

11. Freitas, A.A.: Comprehensible classification models: a position paper. ACM SIGKDD Explor. Newsl. **15**(1), 1–10 (2014)
12. 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)
13. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web **2**(2), 1 (2017)
14. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI-Explainable artificial intelligence. Sci. Robotics **4**(37), eaay7120 (2019)
15. Honegger, M.: Shedding light on black box machine learning algorithms: development of an axiomatic framework to assess the quality of methods that explain individual predictions. arXiv preprint arXiv:1808.05054 (2018)
16. Keane, M.T., Kenny, E.M.: The twin-system approach as one generic solution for XAI: an overview of ANN-CBR twins for explaining deep learning. arXiv preprint arXiv:1905.08069 (2019)
17. Kim, B., Khanna, R., Koyejo, O.O.: Examples are not enough, learn to criticize! criticism for interpretability. In: Advances in Neural Information Processing Systems 29 (2016)
18. Laios, A., et al.: Factors predicting surgical effort using explainable artificial intelligence in advanced stage epithelial ovarian cancer. Cancers **14**(14), 3447 (2022)
19. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. Entropy **23**(1), 18 (2020)
20. Lipton, Z.C.: The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)
21. Lou, Y., Caruana, R., Gehrke, J.: Intelligible models for classification and regression. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 150–158 (2012)
22. Marcinkevičs, R., Vogt, J.E.: Interpretability and explainability: a machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805 (2020)
23. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
24. Molnar, C.: Interpretable machine learning. Lulu.com (2020)
25. Papernot, N., McDaniel, P.: Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. arXiv preprint arXiv:1803.04765 (2018)
26. Rajurkar, S., Verma, N.K.: Developing deep fuzzy network with takagi sugeno fuzzy inference system. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6. IEEE (2017)
27. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
28. Rüping, S., et al.: Learning interpretable models (2006)
29. 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)
30. Zhu, J., Liapis, A., Risi, S., Bidarra, R., Youngblood, G.M.: Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation. In: 2018 IEEE Conference on Computational Intelligence and Games (CIG), pp. 1–8. IEEE (2018)